Resolution-enhanced Digital Epiluminescence Microscopy Using Deep Computational Optics

by

Dino Kabiljagic

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following paper is used in this thesis:

D. Kabiljagic and A. Wong, "Resolution-enhanced digital epiluminescence microscopy using deep computational optics," *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII* (vol. 10881). International Society for Optics and Photonics, 2019.

Abstract

Melanoma is the most common type of cancer, and the standard practice used for examining skin lesions is dermoscopy, where dermatologists use an epiluminescence microscope (ELM) to visualize the skin's surface and subsurface structures for anomalies. Conventional ELM instruments are being replaced by digital ELM instruments that enable dermatologists and other health care practitioners to digitally capture, archive, and analyze skin lesions using computer-aided diagnosis (CAD) software. One of the limiting factors of digital ELMs is a trade-off between spatial resolution and field of view (FOV), where a large FOV, which is needed to allow for larger skin lesions to be examined in their entirety, can be achieved by reducing magnification at the cost of spatial resolution (leading to a loss of fine details that can be indicative of malignancy and disease). In this thesis, we introduced the deep computation optics (DCO) framework for the purpose of resolution-enhanced digital ELM to improve the balance between spatial resolution and FOV. More specifically, the multitude of parameters of a deep computational model for numerically magnifying digital ELM images were learned through a wealth of low-resolution and high-resolution digital ELM image pairs. The proposed DCO approaches were experimentally validated, demonstrating improvements in the spatial resolution of the resolution-enhanced digital ELM when compared to more conventional methods, such as bicubic interpolation. Furthermore, we have demonstrated that the spatial resolution-enhancement improvements can be made within the deep computational models themselves where the model's receptive field is of the utmost importance since the missing information is better estimated when there is a larger number of neighbouring pixels involved.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Alexander Wong. Not only are you fantastic at your job, but you are also a great character. Thank you for your support and enthusiasm.

I would like to thank Prof. David Clausi and Prof. George Shaker for taking the time to read and provide feedback for my work. I greatly appreciate it.

I would like to thank Ivana and Denis for proof reading this thesis. You may have been unfamiliar with the topic, but hopefully you are experts now.

Finally, I would like to thank my wife for the unconditional love and support throughout this journey. Thank you for your patience, and thank you for taking care of our babies when I was busy. Also, thank you majka Nura.

Dedication

I dedicate this thesis to Dina, Ari and Una.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Skin cancer is the most common type of cancer, and according to the World Health Organization (WHO), one in every three cancers diagnosed today is a skin cancer. Furthermore, the American Cancer Society estimates that more than five million new cases of skin cancer will be diagnosed in 2019 in the United States (US) alone, of which more than seven thousand will result in death [3]. In fact, the number of skin cancer diagnoses has increased over the years, and this trend may be due to an aging population, exposure to ultraviolet (UV) radiation as ozone levels are depleted, and greater overall health awareness [3]. Despite this increasing diagnosis trend, the skin cancer survival rate has also increased [3], and this trend could perhaps be attributed to regular cancer screening resulting in early diagnosis and potentially a more positive outcome. Whatever the reasons might be, one observation is evident: just as with any other types of cancer, skin cancer survival rate increases with early detection, and for this reason, it is imperative for dermatologists and health care practitioners to become more knowledgeable regarding the subject in order to be able to quickly assess and diagnose skin cancer.

In order to be highly accurate in their diagnosis, dermatologists rely on a high level of details of the skin's morphological features, such as blood vessels and cellular layers, as well as structure and landscape. For years, the tool of choice for many dermatologists and health care professionals has been dermoscopy via a epiluminescence microscope (ELM) (often referred to as a dermoscope or dermatoscope). This instrument contains a light source, a magnifying lens, and in some cases an optical filter [5], all of which aid in the production of a more detailed visual of the skin's surface, and thus may provide a more accurate diagnosis. [21] revealed that since the invention of ELM and its leveraging for dermoscopy, the accuracy of clinical diagnosis increased to 95 percent, as opposed to the 60 percent accuracy rate of the unaided eye, making it an invaluable tool in early cancer

diagnosis. However, one concern with these analog ELM instruments (Figure 1.1) is that they do not possess a functionality that enables them to record and store images for later viewing as in teledermoscopy [28].

Figure 1.1: Overview of the conventional epiluminescence microscopy. Morphological features of the lesion are examined with the help of magnifying lens with the built-in illumination system and cross-polarizing filter. This analog instrument has no capability to capture and store images for later viewing and consultation.

Digital ELM, on the other hand, enables dermatologists and health care practitioners to

capture and archive digital images of skin lesions. This approach has numerous advantages over conventional ELM given that archived images may be retrieved and further examined by dermatologists with many years of experience. This is particularly valuable to the patients who need systematic treatment for their skin conditions. Furthermore, a study [35] revealed that teleconsultation of clinical and dermoscopic images sent to physicians via e-mail yielded a similar degree of diagnostic accuracy as face-to-face diagnosis, suggesting that teledermoscopy could be further utilized as a triage service aimed at the patients with critical skin conditions. In addition, the use of digital ELM enables clinical decision support to take place using computer aided diagnostic tools [2, 19, 1, 39].

Despite the obvious benefits of digital ELM, there exists a fundamental trade-off between spatial resolution and field-of-view (FOV) that stems from the nature of their purposes. More specifically, since the skin lesions can vary significantly in size (from the tiny dots nearly imperceivable to human eye to the ones covering a part or the entirety of the body), digital ELM instruments have to be designed to ensure that all skin lesions across the size spectrum can be captured in their entirety. In order to account for lesion size variation, a large field-of-view (FOV) is achieved by reducing the magnification at the cost of spatial resolution. This leads to a significant loss of fine details in the image that can be crucial for diagnostic purposes.

1.1 Thesis Scope and Contributions

To improve the balance between FOV and spatial resolution, this thesis introduces deep computation optics (DCO) for the purpose of resolution-enhanced digital ELM. More specifically, the DCO approach in this thesis leverages a deep computational modelling approach based on deep learning [23] to learn the abundance of parameters through the multitude of low-resolution and high-resolution image pairs for numerically magnifying digital ELM images. The proposed DCO approach was experimentally validated, demonstrating improvements in the spatial resolution of resolution-enhanced digital ELM by two-fold while maintaining FOV for skin lesion captured using digital ELM instrument. There are two main contributions:

1. Deep computational optics (DCO) framework based on deep learning outperforms conventional magnification methods used in practice.
2. Spatial resolution-enhancement accuracy depends on network architecture where a large receptive field of the network yields higher accuracy.

The two main contributions in this thesis aim to introduce deep learning method as a valid solution to an ill-posed problem of spatial resolution-enhancement in digital epiluminescence microscopy field. This method can assist dermatologists in increasing of the clinical-decision making accuracy.

1.2 Thesis Overview

The upcoming chapters of this thesis are organized as follows. Chapter 2 reviews all relevant background concepts, including the history and principles of epiluminescence microscopy, followed by digital ELM and current resolution-enhancement methods. The deep computational optics framework is represented in Chapter 3, followed by a brief description of the most important building blocks and concepts used in the framework design. Method implementations, comparison of the test results with a conventional method, as well a comparison between two DCO methods are detailed in Chapter 4. We complete our work by sharing concluding thoughts and future directions in Chapter 5.

Chapter 2

Background

In this chapter we describe a brief history of epiluminescence microscopy followed by a detailed explanation of its principles, from very simple ones utilizing magnification only, to highly complex ones employing optical filters. We also introduce digital epiluminescence microscopy, which is the latest technique employed by dermatologists, and we subsequently describe its spatial-resolution shortcomings and the available methods for tackling such issues.

2.1 History of Epiluminescence Microscopy

The study of the skin's surface dates back to the 1660s, when Borrelus, and subsequently, Kolhaus investigated small vessels in the finger's nail bed using a microscope [44]. Approximately two centuries later, in 1893, German dermatologist Unna, recognized that the upper layers of the epidermis (i.e., the outermost layer of the skin) blocked the light from entering the skin and determined that the skin can be made more translucent in combination with water-soluble oils [44, 5]. In the 1920s, another German dermatologist, Johann Saphier, reported some interesting morphological features of the skin using a binocular microscope with a built-in light source [5]. This skin surface microscopy principal was later picked up and further developed in the U.S. by Goldman during the 1950s [13, 14], when he reported his findings using the device he called "Dermoscopy." He was also the first dermatologist to use this new device to examine the pigmented skin lesions (PSL). Fritsch and Pechlaner utilized the dermoscopy techniques in pre-surgical evaluation of the PSL when they described the pigment network, which is even in present day a widely used criterion for distinguishing between benign and malignant skin lesions [33]. Trying

to improve the early diagnosis of melanoma, Pehamberger et al. in 1987 coined the term epiluminescence microscopy (ELM). They used a surface microscope in combination with oil immersion to analyze a wide range of new morphological features that became apparent with this new technique [34, 43]. Epiluminescence microscopy, dermoscopy, dermatoscopy, and surface microscopy are all names that refer to the same imaging technique of the skin's surface described in the next chapter.

2.2 Principles of Epiluminescence Microscopy

Epiluminescence microscopy, most commonly referred to as dermoscopy, is a non-invasive in vivo examination of the skin lesions that uses a handheld microscope called a dermatoscope (or dermoscope), which, in its simplest form, contains a magnification lens and a light source. This device allows dermatologists and healthcare practitioners to better visualize the skin's morphological features such as size, shape, and colour. However, one aspect of epiluminescence microscopy that makes it an invaluable instrument in dermoscopy is its ability to make structures that are beyond the skin's surface, and therefore invisible to the unaided eye, accessible to the microscopic examination.

2.2.1 Skin-Air-Eye

To understand how ELM provides this additional information, one needs to understand the optical principles in dermoscopy, and more specifically, the interaction of light with the skin. Given that the refractive index of the stratum corneum (i.e., the outermost layer of the skin) is higher than the refractive index of the air [7], most of the incident light is reflected off the surface of the skin back onto the objective (eye's retina). These back-scattered light paths interfere with each other, obscuring the clear visualization of the light that is reflected directly from the deeper layer of the skin [29]. This is depicted in Figure 2.1 where most of the incident light is reflected off the surface of the skin (red line), interfering with the superficial (green line), and penetrating (blue line) light, and therefore preventing clear visualization of the skin's subsurface structures.

2.2.2 Skin-Glass-Eye

To solve the problem of back-scattered light, early clinical dermoscopes employed a non-polarized light sources for illuminating the skin, such as light-emitting diodes, with 10-fold

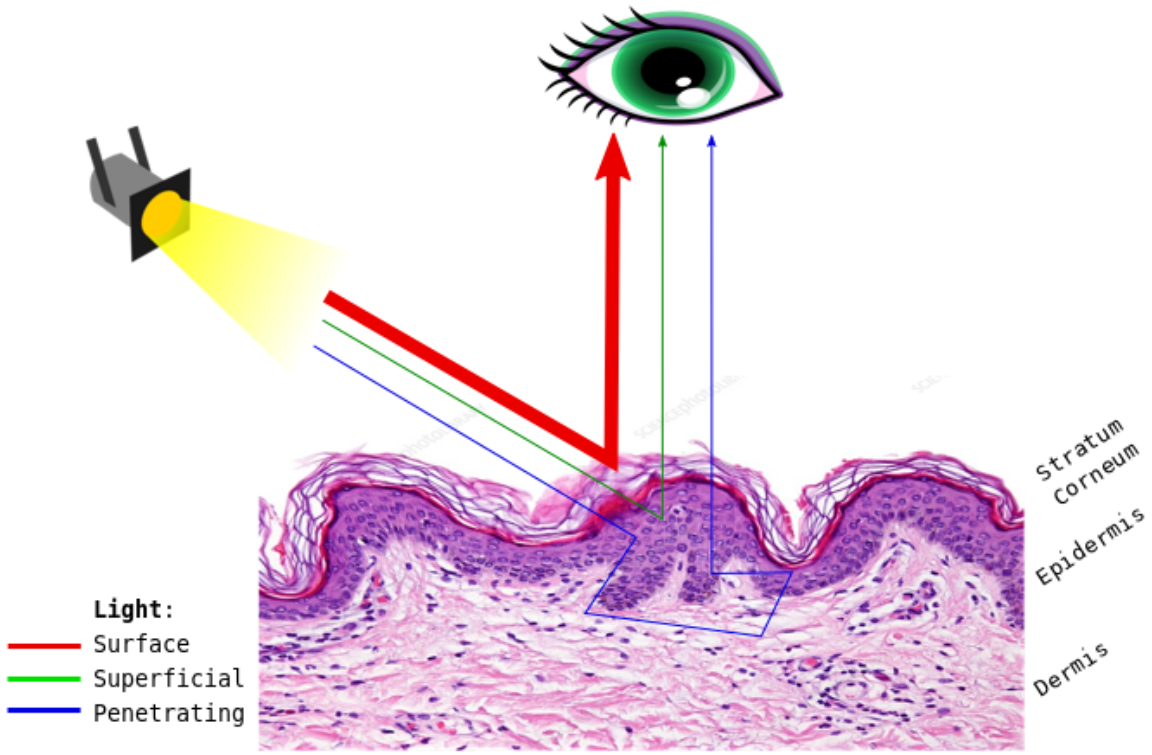

Figure 2.1: Optical properties of the skin. Most of the incident light is reflected back from the stratum corneum (red line), not allowing visualization of the deeper (green and blue lines) skin layers.

magnification lenses. An immersion liquid is applied on the skin's surface and the glass slide is placed on top while applying slight pressure. This skin - liquid - glass setup replaced skin - air setup because the reflective index of liquid and glass is lower than that of air, and therefore light is not reflected but rather absorbed, scattered, and then reflected from structures below the skin surface, allowing the investigators to see through the epidermis [33] (Figure 2.2).

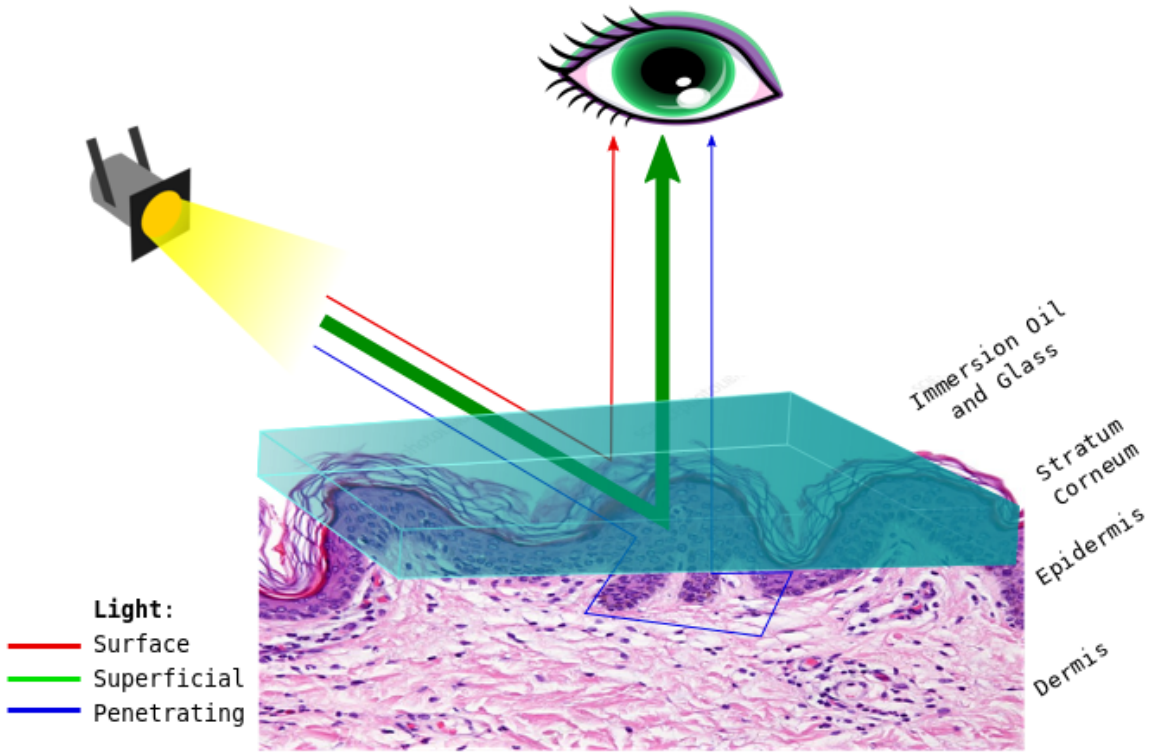

Figure 2.2: ELM with immersion oil and glass eliminates most of the surface glare (red line) allowing for a better visualization of the epidermis (green line).

2.2.3 Skin-Filter-Eye

The next revision of dermoscopes in the year 2000 [29] utilized the principle of cross-polarization. This is a different optical principle than the one described above. Unlike oil immersion and glass setup, the cross-polarized setup requires the light source to be passed through a polarizing filter first, making the incident light unidirectional. This light then passes through another polarized filter which is perpendicular to the first filter, therefore allowing only the light that changes by 90 degrees to reach the objective. This is depicted in Figure 2.3 where the surface light (red line) keeps its original direction and the superficial light (green line) in the epidermis undergoes only minor scattering, insufficient to change by 90 degrees. The cross-polarizier blocks these back-scattered lights. Only the penetrating

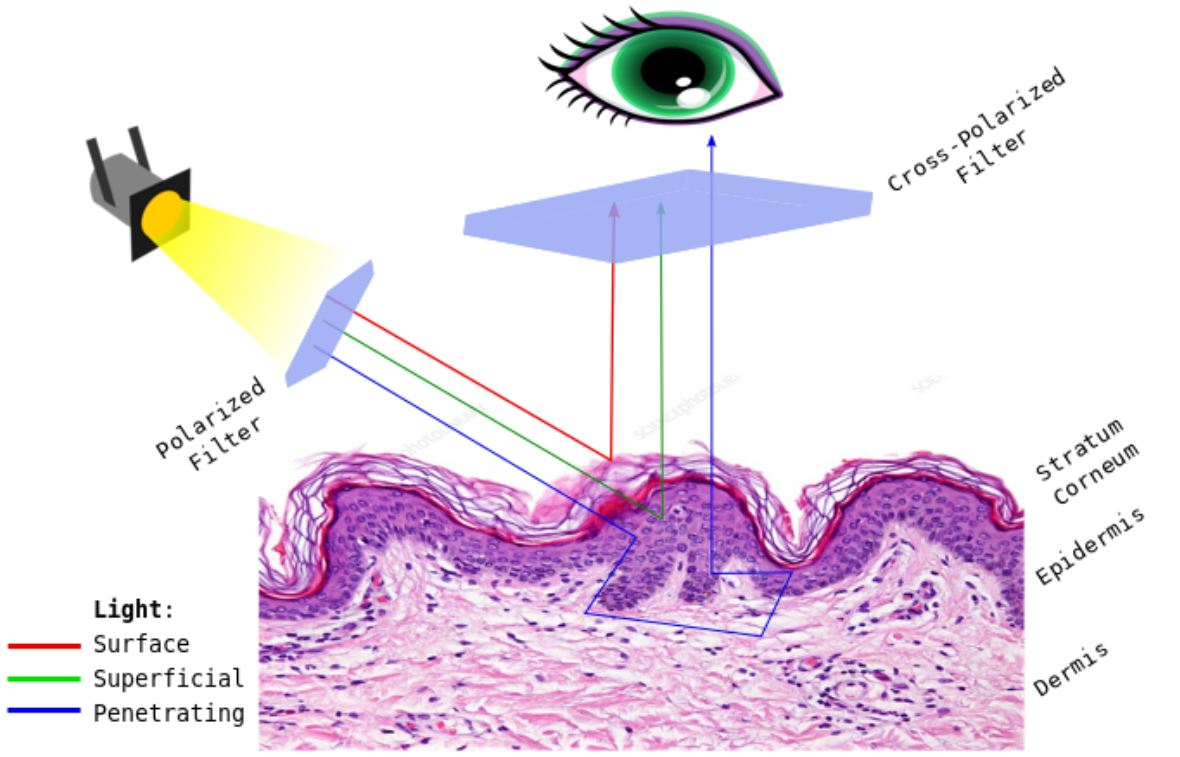

Figure 2.3: ELM with cross-polarized filter blocks the surface and superficial lights, allowing only light that has undergone a number of scattering events in epidermis to be visualized.

light (blue line) undergoes severe scattering in the epidermis, and some of it is changed by 90 degrees, which passes through the cross-polarizer and enables the investigator to examine the dermal properties of the skin.

There is no doubt that the invention of ELM opened up a new era of visual possibilities not attainable with the naked-eye examination. ELM has allowed skin professionals to conduct examinations beyond the surface of the skin and identify new morphological features that may be a critical piece of a puzzle for their clinical decision-making. Furthermore, these new features enabled dermatologists to create universal rules and algorithms for diagnosing malignant and benign tumors. However, one bottleneck of these devices to be even more valuable in their purpose, comes from the fact that they are analog instruments,

and as such do not have the capability to capture and store digital images. But with the recent technological advancements, especially the internet and digital still cameras, there is a novel opportunity to take ELM one step further in their capability, which will be discussed in the next section.

2.3 Digital Epiluminescence Microscopy

Although dermoscopy has been shown to improve clinical diagnostic accuracy of benign versus malignant skin lesions, there exist numerous cases where the lesions do not exhibit specific dermoscopic features, making it almost impossible to diagnose even with the most advanced dermoscopes. However, it has been shown that these "featureless" lesions can be enhanced by observing their changes over time, making it possible for dermatologists to spot the cancer before it becomes an issue [30, 38, 42]. This surveillance of lesions is achieved by capturing digital still images employing ELM in combination with digital cameras. This is called Digital Epiluminescence Microscopy (DELM), or Digital Dermoscopy, and refers to the acquisition and computerized manipulation of dermoscopic images [10].

This new technology enables dermatologists and health care practitioners to capture and archive digital images, which can later be retrieved and further examined individually or in consultation with other skin physicians. In addition, DELM images can be sent instantaneously over the network for real-time consultation, and it enables clinical decision support to take place using computer-aided diagnostic tools. Moreover, this new approach may help reduce the subjective nature of diagnosis, and in combination with clinical rules and algorithms, make it more objective, which will yield a more accurate diagnoses.

When compared to the exponential growth of the technology, digital dermoscopy has and continues to narrow the gap, especially with imaging innovations. Over the years, we have seen digital dermatoscopes of all sizes and shapes, ranging from cart-on-wheel systems, conventional digital cameras, DSLR cameras, and most recently the systems that utilize smartphone cameras, camera clip-on gadgets, as well as systems that utilize Internet-Of-Things (IOT) approach (wireless connectivity and in-cloud computing). Some of these digital dermatoscopes are presented in Figure 2.4.

Despite the obvious benefits of DELM, there exists a fundamental trade-off between spatial resolution and field-of-view (FOV) that stems from the nature of their purpose. More specifically, since skin lesions can vary significantly in size, digital ELM instruments must be designed to cover a wide spectrum of size variations. For example, current DELMs range from 10x10 mm FOV to 40x30 mm FOV for Vivacam by CaliberID and DermLite

Figure 2.4: Digital dermoscopes: (A) Bodystudio by FotoFinder, (B) Portable unit by Vidix, (C) DLF2Plus by Dermlite, (D) VisioMed by Canfield, (E) Vectra 3D by Canfield, (F) Medicam 1000 by FotoFinder, (G) MoleScope by MetaOptima, (H) Handyscope by DermLite

Cam by DermLite, respectively. This large field-of-view is achieved by reducing magnification at the cost of spatial resolution, which leads to a significant loss of detail in the image that may be crucial for diagnostic purposes.

2.4 Current Resolution-enhancement Methods

As soon as it was possible to load digital images onto a computer, researchers have been harvesting the computers' computational power to use signal processing techniques to develop tools for automated analysis of digital images. The end goal of these tools, amongst many, includes object detection, classification, segmentation, and image registration and enhancement, for different application areas such as satellite imaging, security and surveillance, medical imaging, and many more. Image enhancement, and more specifically spatial resolution-enhancement, is one of the oldest and most researched problems in computer vision. It is also one of the most difficult problems to solve due to its ill-posed nature,

where there exist multiple solutions to a single problem, depending on the assumption being made. Also known as digital zooming, or up-sampling, spatial resolution-enhancement involves increasing the size of an image while also increasing the total number of pixels for the purpose of offering more details of the scene, and that could be critical in the decision-making process, especially in dermoscopy. This is depicted in Figure 2.5 where we have a small 3x3 gray-level image in (a), and its 2X magnified (resolution-enhanced) counterpart in (b). Image in (b) has four times the number of pixels than the image in (a), and the idea is to map all the pixels (nine of them) in (a) to all the pixels (thirty six of them) in (b). This is not one-to-one mapping, and for every one pixel in image (a) we have four new pixels in image (b), that need to be estimated, making resolution-enhancement a difficult problem since there are infinitely many possible choices.

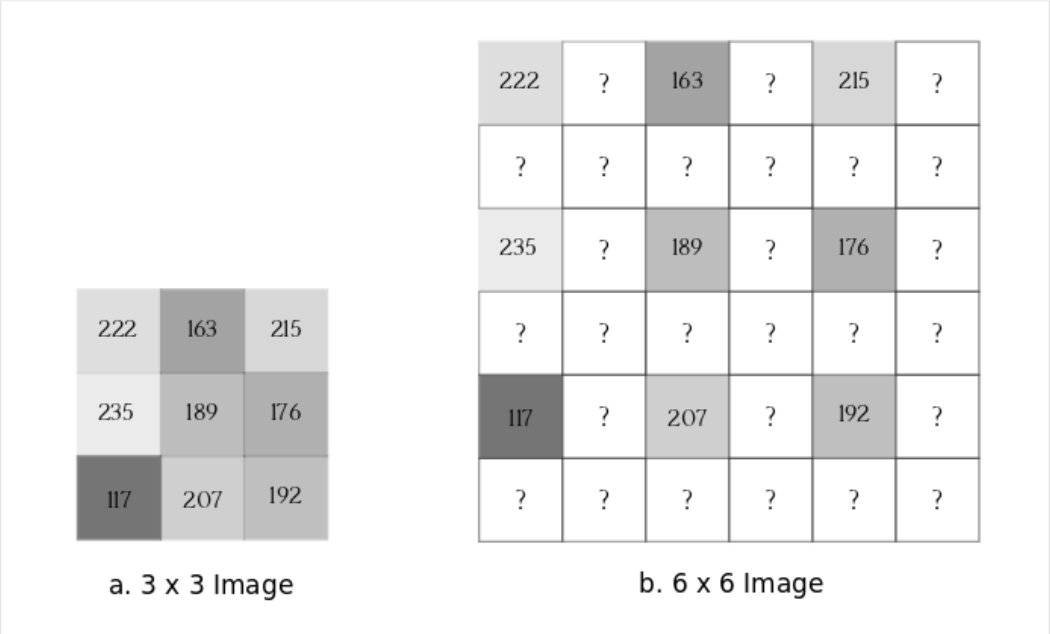

Figure 2.5: Spatial resolution enhancement. (a) Original gray-level image of 3x3 size. (b) 2X magnified version of image in (a). Magnified image in (b) has missing pixels values that need to be estimated based on some assumption.

To date, resolution-enhancement methods can be divided into three main categories: interpolation-based, reconstruction-based, and learning-based methods. All three methods have one goal; the estimation of the value of the missing pixels in the resolution-enhanced image.

Interpolation-based methods such as bicubic interpolation [18] and Lanczos resampling [8] estimate new pixel values based on the known pixel values. Bicubic interpolation, for example, estimates these values by taking a weighted average of the 16 neighbouring pixels, and different pixels are given different weights, depending on their position relative to the central pixel. More specifically, pixels that are closer to the missing pixel are given higher weight in the estimation process than the pixels further away. The bicubic interpolation method is based on direct manipulation of pixels, without considering any content of the image such as pixel intensities, edge information, texture, etc. [32]. As such, they do not preserve high-frequency content, resulting in images with blurry edges and generally suffer from the shortage of accuracy [50]. These methods are simple and fast from the computational standpoint, but since dermoscopy relies on high levels of detail of the skin lesion, interpolation-based methods are not very useful. Despite their shortcomings, the bicubic interpolation method for example is still, forty years later, one of the most widely used resolution-enhancement methods which can be found in most commercially available software packages.

Generally speaking, one way to deal with the inherent ill-posed problem is to constrain the solution space. In that regard, reconstruction-based methods normally preserve high-frequency content by hypothesizing some prior knowledge about the image. For example, [31] uses level-set contour smoothness as a prior to preserve the fidelity of the edges, while [45] uses gradient profile prior, which describes shape and sharpness of the gradient profile to achieve sharp details while preserving the edges and other high-frequency content. In addition, [45] observed that the shape statistics of the gradient profiles in natural images was independent of the image resolution, suggesting a strong correlation of the sharpness of the gradient profiles between HR image and the LR image. This allowed for a constraint on the gradient field of HR image, which in turn resulted in recovery of a high-quality HR image. However, one problem with this approach is that the individual estimation of sharpness for each gradient profile is not robust due to the noise. This is problematic for dermoscopic images that have high-count of high-frequency content, which in addition to edges, exhibit other superficial patterns such as blobs, that consist of globules an dots. Finally, reconstruction-based methods are computationally expensive and there exist fundamental limits when it comes to magnification factors [26].

At the end of the 1990s, learning-based methods, also known as example-based methods, started to dominate the computer vision field due to their fast computation and even higher accuracy than the reconstruction-based methods. These methods lean on machine learning modelling that uses statistical characterization of data to compute decision boundary that yields a high prediction accuracy. Markov Random Field (MRF) was first used for the resolution-enhancement problem by Freeman et al., where a large internal database of low-

and high-resolution image patches is created, and only a fraction of the most similar patch pairs from the database is used to estimate the missing pixels in the resolution-enhanced image [11]. Inspired by sparse signal recovery theory, sparse coding methods [49, 53] use sparse linear combination of high-resolution patches that can be successfully recovered from the down sampled image patches to generate a resolution-enhanced image. Most of these methods used the concept of feature extraction and statistical classifier [27] where the most crucial step in the design of such a system is the extraction of discriminant features. This step is performed by hand (handcrafted features) and lacks robustness and suffers from the so-called "curse of dimensionality," where the computation becomes very expensive due to high dimensional space. In other words, the capacity of the system is limited.

Over the years, we have seen systems designed entirely by humans, to hybrid systems using statistical models and handcrafted features. The next logical step is to design a system that is fully automated, and the idea is to let computers automatically learn features that represent the data optimally. This concept formulates the basis of almost all deep learning models, where the pipeline is composed of many layers and the number of layers is proportional to the capacity of the model to learn higher level features. The ability of the deep network to easily scale up is particularly important to the ill-posed problem of spatial resolution-enhancement, where there are infinitely many possible solutions. In addition to increased capacity, the depth of the network is also directly proportional to the number of non-linear functions, allowing for a design of low-resolution to high-resolution image mapping function that is highly non-linear and therefore highly expressive. This is of paramount importance for dermoscopic images that exhibit high-level of fine details, that previous methods were unsuccessful in reconstructing.

These deep learning methods have recently outperformed the previous methods on multiple benchmarks, from digit recognition, image classification, segmentation, to a resolution enhancement, and many more. In the next chapter, we will describe convolutional neural networks [22], a unique type of machine learning approach, that formulates the basis of our Deep Computational Optics architecture in this thesis.

Chapter 3

Deep Computational Optics

In this chapter we investigate the feasibility of improving the balance between field-of-view and spatial resolution of digital ELM images via deep learning. We first begin with Section 3.1 where we outline some of the simplest optical elements utilized in scene reproduction, and explain how computational optics improve this conventional approach. In section 3.2, we then describe how deep neural networks help in improving computational optics... need to pretty it up

3.1 What is Computational Optics?

Optical systems involve light, which also acts as an input to the sight (the primary sensor for human beings), and these optical systems have been around for centuries to enhance humans' lives. In its simplest form, optical instrument contains a lens and/or a mirror, which helps in a production of a scene, but do not often exceed the capabilities of a human eye [9]. In addition, these basic elements suffer from manufacturing defects, and spherical and chromatic aberrations that appear as a consequence of a nature of light and its interaction with other materials. To resolve these artifacts, and to extend beyond the capabilities of a human eye, optical systems have to be designed to include many basic elements which exponentially increases the cost as well as the size of these systems. However, with the advent of computers, a new paradigm in optics was born, where researchers have leveraged computational power to significantly increase the performance of optical systems consisting of only the traditional optical elements. As such, computational optics enable a design of optical systems that are compact and cost efficient, and often times allow scientists and engineers to extend beyond physical limitations of traditional optical systems

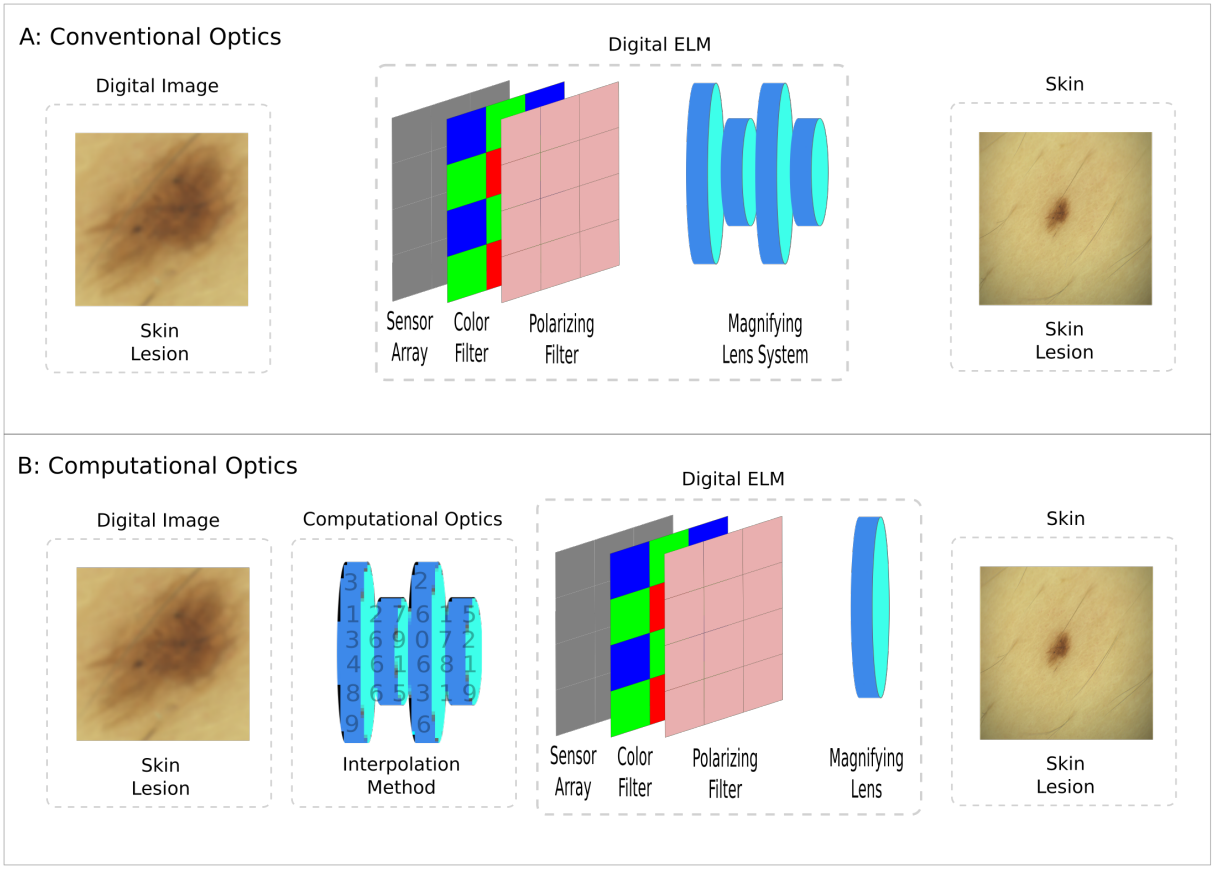

Figure 3.1: (A) The skin lesion under examination may only cover a small portion of FOV of the digital epiluminescence microscope (DELM), resulting in the skin lesion in the digital ELM image to appear low resolution and a significant loss in detail. (B) The addition of deep computational optics (DCO) for the purpose of resolution-enhanced digital ELM enables improved spatial resolution while maintaining FOV.

such as numerical apertures as well as eliminate the need for optical elements in general [4] [55]. Figure 3.1 illustrates a comparison between conventional and computational optics approaches. More specifically, in Fig 3.1(A), conventional optics design incorporates a complex lens system in order to magnify the skin lesion under the examination. On the contrary, computational optics design in Fig 3.1(B) produces the same magnification image of the skin lesion, while using only a single convex thin lens, making the digital ELM lighter in weight, more compact, and less expensive. In their nature, optical elements can be mathematically modelled as nonlinear functions, and one of the most powerful ways to

learn nonlinear function is via deep learning, which we will discuss in the following section.

3.2 Deep Computational Optics via Deep Learning

Despite the the obvious benefits of digital ELM via computational optics such as resolution-enhancement, there are certain limitations to this approach. More specifically, in Figure 3.1(B), the skin lesion under examination only covers a small portion of the FOV of the digital ELM (and thus covering only a small number of pixels in the imaging sensor), resulting in the skin lesion in the captured digital ELM image to appear at low resolution and exhibit a significant loss of detail. This makes clinical screening of such skin lesions using digital ELM more difficult as the missing detail may be crucial to the diagnostic process. By integrating a deep computation optics component into the digital ELM (Figure 3.2), one can enhance spatial resolution of the produced digital ELM image, thus providing greater skin lesion detail for potentially improved clinical diagnosis.

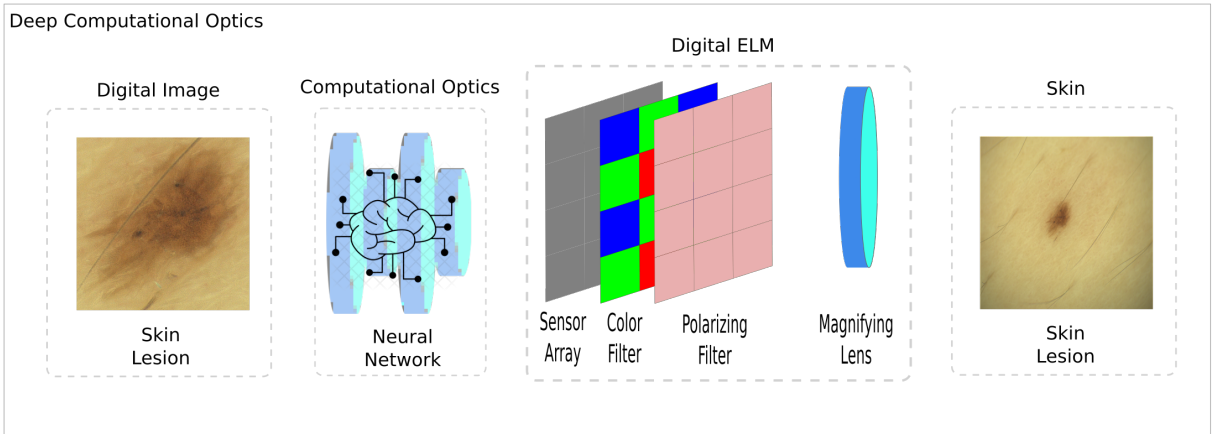

Figure 3.2: The skin lesion under examination may only cover a small portion of FOV of the digital epiluminescence microscope (DELM), resulting in the skin lesion in the digital ELM image to appear low resolution and a significant loss in detail. (B) The addition of deep computational optics (DCO) for the purpose of resolution-enhanced digital ELM enables improved spatial resolution while maintaining FOV.

We propose a deep computational optics component architecture based on convolutional neural network (CNN), one class of the machine learning models that is particularly successful with structured data such as images. More specifically, DCO is based on a

special type of CNN networks, named residual networks, which will be explained in the following sections.

3.2.1 Convolutional Neural Networks

CNNs are very similar to regular artificial neural networks (ANNs) which are made up of neurons that have learnable parameters. However, neurons in CNNs are only connected to a small region of neurons in the previous layer, as opposed to every single neuron like ANNs. Even though they have been around for decades [24], it was not until recently that they gained a wide audience interest, particularly, when their deep architecture won the ImageNet LSVRC-2010 competition by a large margin compared to a runner-up [22]. Since this win, deep CNN architectures have been successfully applied in high-level computer vision tasks, such as classification, object detection, segmentation, and many more. However, since spatial resolution-enhancement is regarded as a low-level computer vision problem due to the pixel-level processing, it is important to understand the ins and outs of CNNs in order to apply them to low-level problems like resolution enhancement, inpainting, deblurring, denoising, super-resolution, and so on. One would expect that some high-level computer vision architectural parts of the CNN would be applicable to low-level computer vision problems, while others may not. For that reason, the next few sections describe the main building blocks of CNN and their functions, as well as some concepts specific to resolution enhancement.

Convolutional Layer

As a high-level overview, CNNs are made up of sequences of layers, and each layer takes a 3D volume as an input and transforms it into a 3D volume output. This transformation from the input to the output is executed by a **convolutional layer**, which is also the most important building block in the CNN architecture. The parameters of the convolutional layer consist of learnable filters, which can vary in size but are always symmetrical. Furthermore, filter sizes are almost always odd numbers like 3x3, 5x5, 7x7, etc., and each convolutional layer will have a specified number of filters, that will be responsible for picking out and detecting patterns. Each filter is convolved over a block of pixels of the same size as a filter in the input volume (i.e., the receptive field of the filter), and produces an output we call the activation map. This is depicted in Figure 3.3 below. Intuitively, the network will learn filters that learn patterns (or features) like edges, corners, blobs, and as we progress deeper into the architecture, filters learn more complex features. When it

comes to the size of the filter, Simonyan and Zisserman [41] showed that a smaller receptive field with multiple layers is as effective as the large filter size, but with many more parameters. Based on this intuition, we will only work with kernels of the 3x3 size.

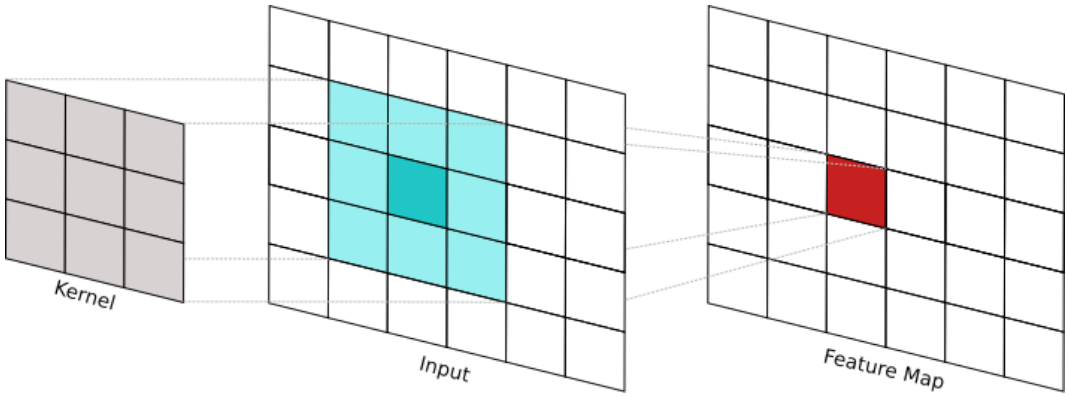

Figure 3.3: Illustration of the convolutional layer. Filter is convolved over a block of pixels in the input volume, and produces an output named Feature Map.

Convolutional layer contains parameters and hyperparameters that need to be carefully selected when designing an architecture. As mentioned earlier, each filter in ConvNet contains weights and a bias, and the number of filters in a single layer is a hyperparameter named **depth**. This is different from the network depth, which refers to the number of layers in the architecture. Each of these filters in a layer learns something different about the input volume, and this capacity of the convolutional layer to learn is one of the key reasons why the ConvNets have been so successful in high-level computer vision tasks. However, one should be careful when choosing the number of filters in the layer, as they come at the price of computational overhead.

Another very important hyperparameter in a convolutional layer is **stride**, which is a fixed number and specifies how many pixels we move the convolving filter. A stride of 2 means the filter skips every other pixel in the input volume which produces an output volume that is smaller than the input volume. This is similar to spatial down sampling, which is the opposite of our spatial resolution-enhancement task, and for that reason, it will be left out as a possible option here. A stride of more than two has the same affect resulting in downsampling and is very uncommon in practice, therefore we are left to work with a stride of one only. The last but not least hyperparameter in a convolution layer is **zero-padding**. Each time an input volume is convolved with a kernel, the output volume is smaller. Due to the convolution operation, the border of the input volume is reduced symmetrically. To keep the input and output volumes the same, we would like to pad the

borders of the input volume with zeros in such a way that the output volume is the same as the input.

Activation Function

In a CNN, the convolutional layer is typically followed by an **activation function**, which is a thresholding non-linear transformation over the input. It helps decide if the dot product between the filter and the block of pixels in the input volume should or should not contribute to the output volume. In other words, it helps determine whether a neuron should fire or not. There are a few different activation functions such as sigmoid, tanh, and ReLU, and each has its advantages and disadvantages. For example, one issue with sigmoid and tanh is that they force large values to equate to one and small values to equate to negative one and zero for tanh and sigmoid, respectively. In other words, these functions are only really sensitive around mid-points. This may pose a challenge for the model to make further improvements as the learning continues, limiting the model's capacity to learn new complex features. ReLU or Linear Rectified Unit, on the other hand, is the most popular activation function to date. Mathematically, it is defined as $y(x) = \max(0, x)$, which means it is linear for all positive values, and zero for all negative values. Unlike the previous two activation functions that are expensive to compute, ReLU is not, since we threshold the activations at zero. Furthermore, its linear shape means the large values do not plateau like tanh and sigmoid functions, making it converge faster.

With the rapid development of deep learning techniques in recent years, deep learning methods have been actively explored, and there are many different possible combinations of sets of deep learning components such as model framework, network design, learning strategies, and so on. Machine learning researchers integrate the above components into a system that models a specific purpose. This thesis' focus is on spatial resolution-enhancement. Therefore, our deep computational optics component will be limited to using convolutional layer and activation functions, along with deep network strategies that were proven to work with low-level computer vision problems.

3.2.2 Receptive Field of the Network

Generally speaking, images are made up of many spatial frequencies (i.e., scales) ranging from very coarse to medium to very fine. In other words, a naturally occurring scene in an image will contain low frequency areas where the pixels are constant and not changing, like clear blue skies, and high frequency areas where the pixel intensities vary, like trees

and bushes in the image, and leaves on the trees. Similarly, images generated using digital ELM imaging modality will contain a lesion and surrounding skin. The pixel variation of the surrounding skin is likely to be uniform and therefore treated as a low frequency component of the image. In contrast, pixel variation of the lesion will be medium to high, depending on the complexity and nature of the lesion under the examination. Normally, the human visual system is not sensitive to differentiation between different scales for it blends all the frequencies into one visual representation. However, using computer-aided techniques, we can decompose complex images into its constituent spatial frequencies.

Digital ELM instruments reveal the entire spectrum of spatial frequencies, and it is the dermatologist's goal to determine what information within this spectrum is paramount. Typically, the most crucial information resides in the high end of the frequency spectrum, and it is the goal of this thesis to develop an algorithm to make those missing high frequency components available for doctors to support their clinical decision-making. To achieve this goal, we are using deep neural networks and the notion of the "receptive field." A receptive field can be defined as a region in the input volume that a particular feature map is being affected by. In other words, the receptive field of the layer is the actual size of the convolving kernel that is responsible for picking out and detecting discernible features. However, not all pixels in the receptive field equally contribute to the feature. Within a receptive field, pixels closer to the center are more important contributors for the calculation of the feature as opposed to the pixels further away (i.e., pixels at the edge). Therefore, the receptive field not only looks at a certain region, but it also focuses on the center portion of the region. To some extent, this is analogous to the bicubic interpolation method, where the central pixel is calculated using 4x4 neighbourhood (i.e., its effective field of view), and weight distribution varies from pixel to pixel. That is, pixels closer to the center will have bigger weights and as such, will have greater contribution to the final calculation.

As the depth of the network increases by adding more convolutional layers, the overall network's receptive field is increased and can be expressed using $\mathbf{(2D + 1)\ x\ (2D + 1)}$, where D is the number of layers in the network. As we can see, using a filter size of 3 x 3, we need two layers in the network to achieve bicubic interpolation's "receptive field." Therefore, the logical question is, how deep can we make the network and does it make a difference for resolution-enhancement? It turns out that deeper networks are actually preferred, and most modern architectures use this approach, which we will discuss in the next section.

3.2.3 Deep CNNs

Deep neural networks, and more specifically deep convolutional neural networks, can be defined as networks with many hidden layers. These networks have opened a new research area in the artificial intelligence domain because of their unprecedented success in many high-level computer vision tasks from object detection, classification, and segmentation, to name a few, as well as the low-level computer vision tasks, such as image reconstruction and enhancement. This success of deep CNNs stems from the recent evidence suggesting that the deeper models have the ability to integrate all levels of features in an end-to-end fashion, and these features can be complemented by stacking more layers in the network (depth of the network). This is particularly important for the resolution enhancement problem at hand since deeper networks utilize larger contextual information in the image. We can use this information to estimate the new pixels' missing values since collecting and analyzing more neighbouring pixels gives us more clues. In addition, due to the nature of CNNs where an activation function follows convolutional layer, increasing the number of hidden layers increases the number of non-linearities, giving these networks the ability to capture some very complex functions. As a result, we have a very expressive network.

However, despite the positive outlook of the deeper network to gain higher accuracy by simply increasing the number of hidden layers, it turns out that there are some issues with this approach. [12] observed that standard gradient descent in combination with random initialization did poorly in deep networks, and more specifically the activation function used (sigmoid) caused hidden layers to saturate, the so-called vanishing gradient problem. This issue was addressed by "normalized initialization," which helps the network with many hidden layers to start converging by maintaining activation variances. Secondly, as the network depth increases and starts converging by means of normalized initialization, one would expect the accuracy to get better. However, [16] observed that instead of accuracy increasing, it saturates and starts to eventually degrade, leading to a higher error rate during training. To tackle this issue of degradation, [16] developed a deep residual learning framework where they recognized two points: one, instead of learning a function in an end-to-end manner, they only learn residual mapping, and second, they use identity mapping realized by shortcut connections that copies input of the residual block and adds it to the output which is the residual. This will be further explained in the next chapter.

3.2.4 Residual Learning

Traditional Neural Networks take on a sequential and hierarchical approach where each layer feeds into the next. This pipeline assumes that higher abstractions can be created

by adding more layers in the network which results in more complex representations. For instance, the first few layers may recognize simple lines and dots. Then the next few layers may recognize more complex lines and corners. Eventually the deeper layers would be able to pick up much more complex features such as eyes and ears, or in our case morphological features, such as blood vessels and cellular layers, as well as structure and landscape, that are otherwise crucial for dermatologists in their diagnosis.

However, one problem with this approach is perhaps the assumption that each layer is dependent on the previous one only. What if the current layer's accuracy could be fine-tuned by using more than just one input? A unique neural network, named residual network, does just this where information from a layer prior also considers the information from the layer a couple of hops prior. This new piece of information from the skipped prior layers is delivered to the current layer via skipped connection. This idea was proposed by [16] and is depicted in Figure 3.4 where we have a representation of a residual network, and more specifically a representation of a residual block. The residual block's input passes through a convolutional layer, followed by an activation function, and then another convolutional layer. The block's output is residual, that is then added to the input that is delivered via shortcut connection. The addition of the input and output of the residual block becomes the input to the next layer in the network's pipeline.

This new architecture enables the network to preserve information within a larger sample of layers in a network since the input to the block is not discarded but rather added at the output. Furthermore, the shortcut connection that copies input and carries it forward to the block's output allows the network to learn identity function by simply setting residuals, that is, parameters within the block, to zero. This makes the network more dynamic in nature, since, in reality, we do not know the optimal number of layers needed to achieve good results. We allow the network to skip the training for the layers whose residuals are very small or close to zero and do not contribute to the overall improvement.

Chapter 3.2 outlines how images are made up of many spatial frequencies and how every image can be decomposed into low (low-resolution) and high-frequency (high-resolution) counterparts. Furthermore, low- and high-resolution image pairs essentially share the same low-frequency content, and it is this high correlation between the two that allows the network to learn residual's only (high-frequency content), rather than the whole end-to-end mapping. Once the difference between images has been learned, the result is combined with a low-resolution image to produce the final high-resolution image. This can be expressed mathematically by looking at the pipeline of a traditional network; we can see that the network is learning the actual output F(x), given the input x. This can be seen in Figure 3.4 where we can ignore the addition operator and the shortcut connection. However, since we have a large correlation between the input (low-resolution) and output (high-resolution)

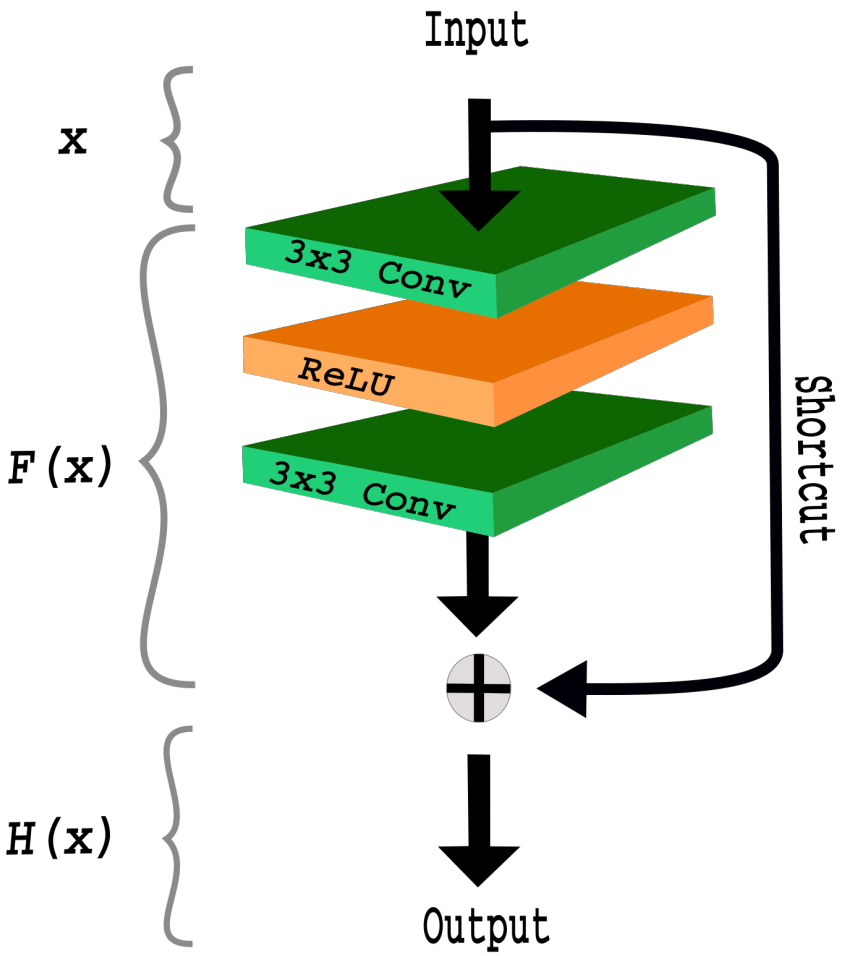

Figure 3.4: Residual block: input x is fed forward through the block, as well as added directly to the output of the block via a shortcut connection.

image, we can define a residual image, H(x), as a difference between output and input, and this residual image will have minimal values and in some instances will be close to zero.

$$Residual = Output - Input$$

$$F(x) = H(x) - x \tag{3.1}$$

Rearranging the above formula, we get:

$$H(x) = F(x) + x \tag{3.2}$$

Like the traditional block that is learning the actual output, the residual network is learning the residuals. These residuals are nothing but high frequency content, or high-resolution detail in the image. For this reason, it makes sense to use this type of neural network for the spatial-resolution enhancement. In the next two sections, we will introduce two different variations of residual networks; VDSR [20] and EDSR [25], respectively.

VDSR inspired DCO

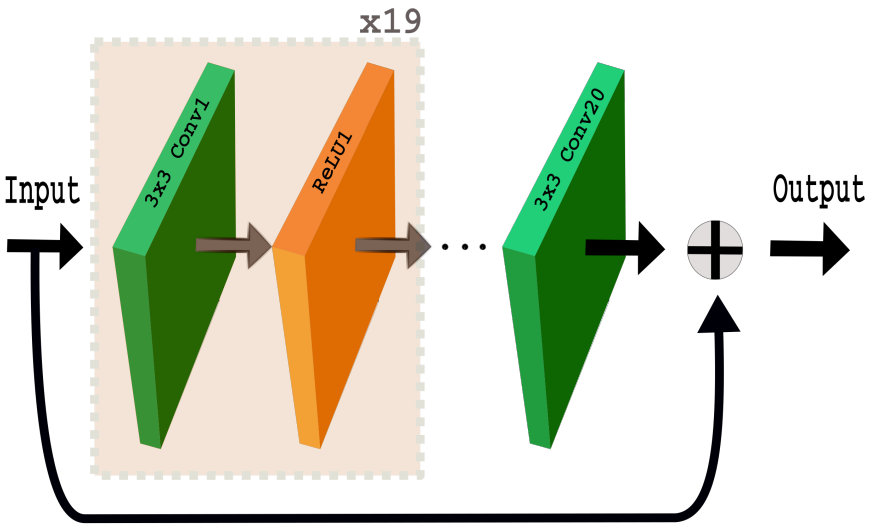

Figure 3.5: Deep computational optics architecture inspired by VDSR network.

The first deep computational optics component architecture in this thesis is based on very deep residual convolutional architecture proposed by [20]. More specifically, as shown

in Figure 3.5, the deep computational optics component architecture consists of a very deep stack of 20 convolutional layers (with ReLU activations, except the last convolutional layer), designed for the purpose of deep residual representation, followed by an element-wise add operator between the stack's output and identity shortcut connection output feeding from the input itself to produce the final numerically magnified image. The input to the network is a low resolution image upsampled 3X using bicubic interpolation. The upsampled low resolution image is carried throughout the network in the form of a feature map, and to maintain the same spatial resolution as the output image, a zero-padding is performed after each convolution. As mentioned earlier, zero-padding can have an artifact around the borders in the final image known as border effect, however, this is an acceptable artifact since we are mainly interested in recovering a lesion that resides in the central part of the image. All of the layers in VDSR, except the first and the last layer, have the same configuration of 64 kernels with each kernel being of 3x3 size. The last layer in the stack is a convolutional layer, that produces a final feature map (residual image) that contains high-frequency content. This image is then added to the original low-resolution image to create a final high-resolution image.

By looking at the VDSR architecure, one can notice the similarity with residual block in Figure 3.4, where instead of a complex residual block containing nineteen ReLU activations, we have the simplest residual block that only contains one activation function. Therefore, one can look at the VDSR as the simplest form of residual network with only one hop. However, the residual function in this network can be seen as the most complex for the residual network, with nineteen activations. As such, we can treat VDSR as a global residual learning since it learns a residual between two images in an end-to-end manner. In addition, since most residuals are zero or close to zero, the complexity of the model as well as the learning difficulty are greatly reduced.

Since spatial resolution enhancement involves increasing the number of pixels in the final image, every new pixel needs to be estimated based on the statistical information about its spatial neighbourhood. In ConvNets, this is achieved using a concept of receptive field of the layer, as discussed in Section 3.2. Rather than using a large receptive field such as 11x11 [22], or 7x7 [51], VDSR uses a very small receptive field of 3x3 cascaded multiple times. It is easy to show that cascading two such small filters is equivalent to a receptive field of 5x5, and the three small cascading filters is equivalent to a receptive field of 7x7. This follows a general formula:

$$Network'sReceptiveField = (2D + 1)x(2D + 1) \tag{3.3}$$

where D represents the number of cascades, which is essentially the depth of the net-

work. Since the filter size is small, it is feasible to have a very deep network without running into the issue of having too many parameters resulting in a computationally expensive system. The advantages of this approach are firstly that the number of total parameters is significantly reduced, and secondly, at each cascade there is a non-linear rectification operator that makes the whole pipeline more expressive [20]. Given the fact that VDSR has 20 layers, the effective field of view at the network level is 41x41. This means that the network uses a large contextual information source (41x41 pixels) to estimate the new pixel value for resolution enhancement problem. This is highly desirable for the spatial resolution enhancement problem for digital epiluminescence microscopy images that require a high level of detail of the skin's morphological features, such as blood vessels and cellular layers, as well as structure and landscape.

EDSR inspired DCO

The second deep computation optics component architecture in this thesis is based on a more complex residual convolutional architecture proposed by [25]. As shown in Figure 3.6, the deep computational optics component architecture consists of a convolutional layer whose output is fed to a very deep stack of 32 residual blocks linearly feeding into each other, followed by one more convolutional layer and then an element-wise add operator that adds the input identity function via skipped connection to the output of the convolutional layer after the deep residual stack. In addition to having a more complex architecture than VDSR, EDSR also has a different input. In fact, input to the network is the original low-resolution image without up-sampling interpolation. As such, the spatial resolution of the feature map is three times smaller than that of VDSR architecture, and since the output of the network has to be of the same size as the ground truth image, feature map is up-sampled at the last layers in the network. Instead of increasing the resolution of the feature map via deconvolutional layer [52] that incorporates arbitrary interpolation method which makes use of redundant pixels in the feature map, EDSR architecture utilizes an effective sub-pixel convolutional layer named ESPCN [40]. Essentially, ESPCN is an array of up-scaling kernels that are specifically trained for every low-resolution to high-resolution map, while it adds minimal computational requirements and in return enables the expansion of the network's depth. EDSR too uses zero-padding to preserve spatial resolution after convolution operator, and each convolutional layer consists of 64 kernels, where each kernels has 3x3 size.

If VDSR architecture can be treated as a global residual learning, one can look at EDSR as a local residual learning where within one main hop, there are 32 local hops and each hop entails two convolutions and a non-linearity in between. That is, VDSR architecture is the

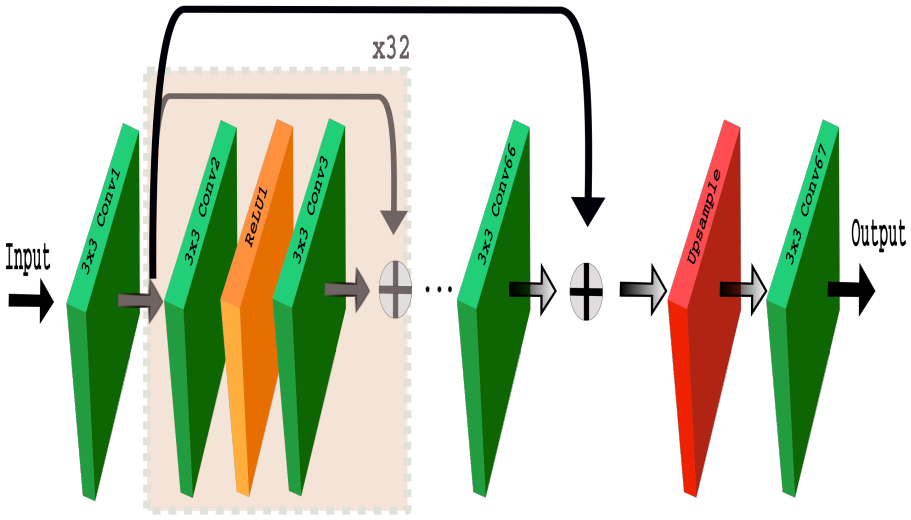

Figure 3.6: Deep computational optics architecture inspired by EDSR network.

simplest form of the residual learning that has one complex residual function, while EDSR is the most complex form of residual learning that has many simple residual functions. Since each residual function adds low-resolution feature map to a high-resolution feature map, the fine image details are learned gradually throughout the network. Just like VDSR, EDSR architecture uses a small receptive field of 3x3 kernel and following the Formula 3.3, the effective receptive field of the network is 135x135, since there are 67 convolutional layers in the network. This is a much larger receptive field than that of VDSR, and in return, we expect much larger contextual information to be included in the reconstruction of the new pixel, resulting in a more accurate prediction of the missing pixel. Large number of features in a neural network such as EDSR usually yields an unstable training [46], and EDSR solves this problem by residual scaling of the last convolutional layer in every residual block. Lastly, it is the combination of large number of layers that increase the networks capacity, low-resolution feature maps that decrease the memory requirements, as well as the residual scaling that stabilizes the training, that make EDSR architecture one of the best deep learning frameworks for the resolution enhancement problem.

Chapter 4

Realization and Experiments

In this chapter we use an empirical approach to evaluate the proposed algorithms. In Section 4.1, we provide dataset information, as well as the low-resolution and high-resolution image pair generation approach. In Section 4.2, we explain metrics used for both algorithm training as well as quantitative results. Furthermore, in Section 4.3, we describe the implementation details of the proposed algorithms, and we conclude with the results in Section 4.4.

4.1 Dataset

The dataset used in this thesis was drawn from the International Skin Imaging Collaboration (ISIC), which hosts the most extensive collection (i.e., over 30 000) of dermoscopic images. More specifically, our data was extracted from the "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge dataset [6][47]. Generally, most of these images contain skin in the background and the mole or the lesion in the foreground, with a few exceptions where the entire image is covered in lesion only. Since this dataset contains images obtained using various dermoscopy techniques, some images are very clear high definition and some contain different artifacts such as blur, and some even have lighting conditions where image backgrounds exhibit unnatural skin colours. As such, these images exhibit relatively low variations, and one would expect the training error to reach optimal values within the first few epochs. Most variations come from the centrally located mole or lesion which is desirable in ConvNet type of networks where some networks generate an undesirable artifact known as a border effect. This artifact is caused by zero-padding after the convolution operator to maintain the same image size throughout the network. All of

the previously mentioned facts make this dataset ideal for the deep learning approach as a possible solution for the spatial resolution enhancement problem.

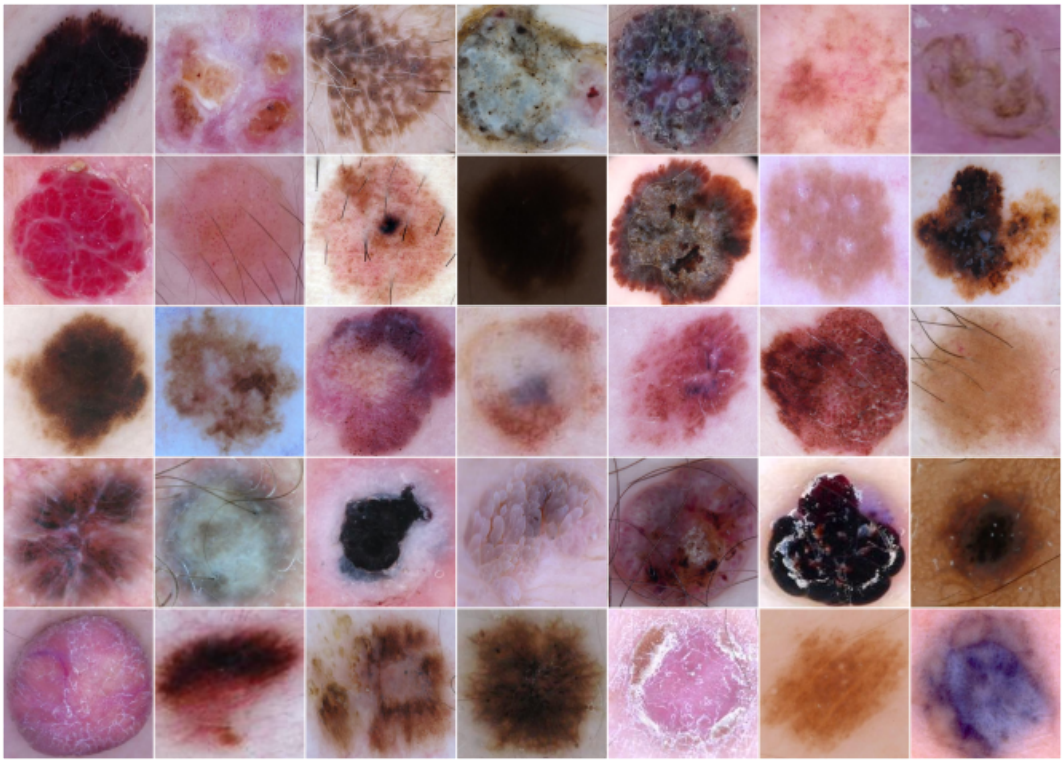

Figure 4.1: ISIC dataset of skin images that contain skin in the background and mole in the foreground.

The essence of supervised learning lies in the data that comes in a format of input data and the label or ground truth. However, ISIC dataset only contains ground truth images, and we have to generate the input data that will be used in the algorithm training step. We depict this data generation process in Figure 4.2, where we take a high-resolution image and create its low-resolution counterpart by downscaling it using bicubic interpolation with a certain level of Gaussian noise and blur. This downscaled version of the original ISIC dataset becomes the input for the training of DCO methods.

Figure 4.2: Dataset generation process of high-resolution and low-resolution image pair for the DCO training. The high-resolution image is downsampled with blur and noise to generate a low-resolution counterpart.

4.2 Metrics

There are two approaches to assessing the quality of digital images: the subjective and objective approaches, or qualitative and quantitative approaches, respectively. The subjective approach involves a human decision-making process based on the perceiver's impression of the quality of the two images. This approach is generally costly in terms of time and budget. Furthermore, it is prone to human error since a person may favour one image at one time and favour a completely different image at a different time. As such, it is rarely adopted in practice. However, we will apply it in this thesis on a smaller data sample when evaluating our results. On the other hand, the objective or quantitative approach is based on mathematical derivation, and is used extensively in the computer vision field for its implementation simplicity. It is important to note that better objective results do not always translate to better qualitative results and vice versa.

The next step after data preparation in a deep learning approach is to assess the quality of the image generated by the network. This quantitative task is achieved by the loss function, often called objective or cost function. The purpose of the loss function is to measure the difference between the network's output and the ground truth image. This comparison is achieved by calculating the difference between the pixels in two images, and for this reason, both images are required to be of the same size and contain the same spatial resolution. The difference between a network generated image and the ground truth image is then sent back to the network, where the parameters are adjusted to generate features

that produce the output image, which will reduce the loss function in the next training iteration. These steps represent the fundamentals of network training.

Most of the quantitative methods are based on two derivations: an absolute value of a difference, and the square of a difference of pixels between two images. Since differences can be positive and negative, and hence their summation can lead to a zero value, to avoid cancellation of values, it is common practice to use absolute value or square of the difference, which we will explain in the next few paragraphs.

By far, the most popular image quality assessment method is mean squared error (MSE) defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{I}_i - I_i)^2 \tag{4.1}$$

where n represents the total number of pixels in an image, $\hat{I}_i$ is the network's predicted image, and I_i is the ground truth image. Also called L2 loss, it computes the average value of the sum of all squared differences of pixel values between two images. In its nature, this method penalizes larger errors since the difference is squared, and it is more tolerant of smaller errors. As such, it is often used when attempting to find an outlier in the dataset. Furthermore, the cancellation of values in MSE formula is handled by $(\hat{I}_i - I_i)^2$ term, which always produces a positive value.

The second quantitative method that is of interest in this thesis is mean absolute error (MAE), defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{I}_i - I_i\right| \tag{4.2}$$

where n represents the total number of pixels in an image, $\hat{I}_i$ is the network's predicted image, and I_i is the ground truth image. Also called L1 loss, it computes the average value of the sum of all absolute difference of pixel values between two images. Unlike L2 loss, L1 loss will penalize smaller values and is more tolerant of the larger values since it does not square the error term. As such, it is less likely to inflate the convergence, which was also reported in practice [54]. Furthermore, just as MSE uses the square term to avoid cancellation of values, MAE uses $\left|\hat{I}_i - I_i\right|$ term to do the same, where absolute value ensures the difference is always positive.

MSE and MAE are excellent metrics in the optimization problem, however, they are not particularly meaningful to researchers without extra information about the image. For

example, an MSE value of 100 is not the same for an image encoded by 8-bits, and an image encoded by 12-bits, and so on. In the field of image processing, the most popular measure when reporting the quantitative results about the image quality is Peak Signal-to-Noise Ratio (PSNR) expressed as:

$$PSNR = 10log_{10}\frac{L^2}{MSE} \tag{4.3}$$

where L is the dynamic range of the image, that is the maximum value of the pixel. For example, images that have 8-bit pixel resolution, $L = 2^8 - 1 = 255$ for uint8 data type, or $L = 1$ for floating point data type (0 to 1), etc. Since PSNR uses MSE as a denominator term in Formula 4.3, they are inversely proportional, and a lower value of MSE will generate a higher value of PSNR. Therefore, using PSNR as a quantitative measure of the image enhancement quality task is not desirable since it will favour MSE over MAE as a loss function.

The second most used quantitative image quality measure in literature is SSIM or Structural Similarity Index Measure [48]. It is a measure of three local elements of the image: luminance (i.e., image brightness), contrast, and structure between two images, and can be expressed as:

$$SSIM(i, \hat{i}) = (\frac{2\mu_i\mu_{\hat{i}} + C1}{\mu_i^2 + \mu_{\hat{i}}^2 + C1})(\frac{2\sigma_i\sigma_{\hat{i}} + C2}{\sigma_i^2 + \sigma_{\hat{i}}^2 + C2})(\frac{\sigma_{i\hat{i}} + C3}{\sigma_i\sigma_{\hat{i}} + C3}) \tag{4.4}$$

where μ_i and $\mu_{\hat{i}}$ represent the means of the local patch, σ_i and $\sigma_{\hat{i}}$ represent the standard deviation of the local patch and the $\sigma_{i\hat{i}}$ is the cross-correlation of i and $\hat{i}$. The terms $C1$, $C2$, and $C3$ are $<< 1$ and serve as a numerical stability. In practice, we usually assess the quality of the entire image, and not just the patch. Hence, we use mean SSIM (MSSIM) which, is the average of sum of all SSIM and is expressed as:

$$MSSIM(I, \hat{I}) = \frac{1}{M}\sum_{k=1}^{M} SSIM(i_k, \hat{i}_k) \tag{4.5}$$

where I and $\hat{I}$ represent original and reconstructed images, respectively, and i and $\hat{i}$ represent original and reconstructed patches, respectively.

4.3 Implementation Details

In this thesis we used Ubuntu 18.04 operating system running on AMD Ryzen 5 2600 processor with 16GB of RAM paired with Nvidia's GeForce RTX 2060 with 6GB of RAM. Deep learning framework used was Pytorch 1.6.

4.3.1 Data Preprocessing

The ISIC dataset set used in this thesis contains over 15,000 images of skin lesions that range in size from a couple of kilobytes (KBs) to over 70 KBs. Since all the training images need to be of the same size (width and height), we extracted 35,000 of the 180x180 non-overlapping patches, which were further divided into 30,000 images for training set, and 5,000 images for testing set. We have also created a smaller rapid-prototyping dataset containing 5,000 of 180x180 patches for training set, and 1,000 patches for testing set. The smaller dataset allowed us to complete additional testing needed for making decisions when it came to certain training options. For example, training the algorithm with and without data normalization revealed that the ISIC dataset also greatly benefits from data normalization when it pertains to training stability and convergence. Data normalization was achieved by computing the entire dataset's average mean and standard deviation for each channel (red, green, and blue) in the image.

4.3.2 Deep Computational Optics I

Generally speaking, deep learning methods benefit from the large training dataset, and as described in the previous section, we trained our DCO I network using one dataset containing 30,000 images and one dataset containing 5,000 images. The test convergence curves for two training datasets are shown in Figure 4.3. Using the same number of epochs, the larger dataset achieves an average PSNR of 32.75dB, which is higher than the PSNR of 31.88dB achieved by the smaller dataset. The convergence test curves for two datasets behave in a symmetrical pattern, suggesting that the training is consistent throughout the whole training process. Additionally, training error for both datasets reaches optimal values after 20 epochs, which is indicative of low variance data, as suggested in Section 4.1.

The pipeline of our first algorithm entailed taking a high-resolution image and creating its low-resolution counterpart for training. This was achieved by first down-scaling the HR image by a factor of three, and then upscaling it by the same factor of three. This way,

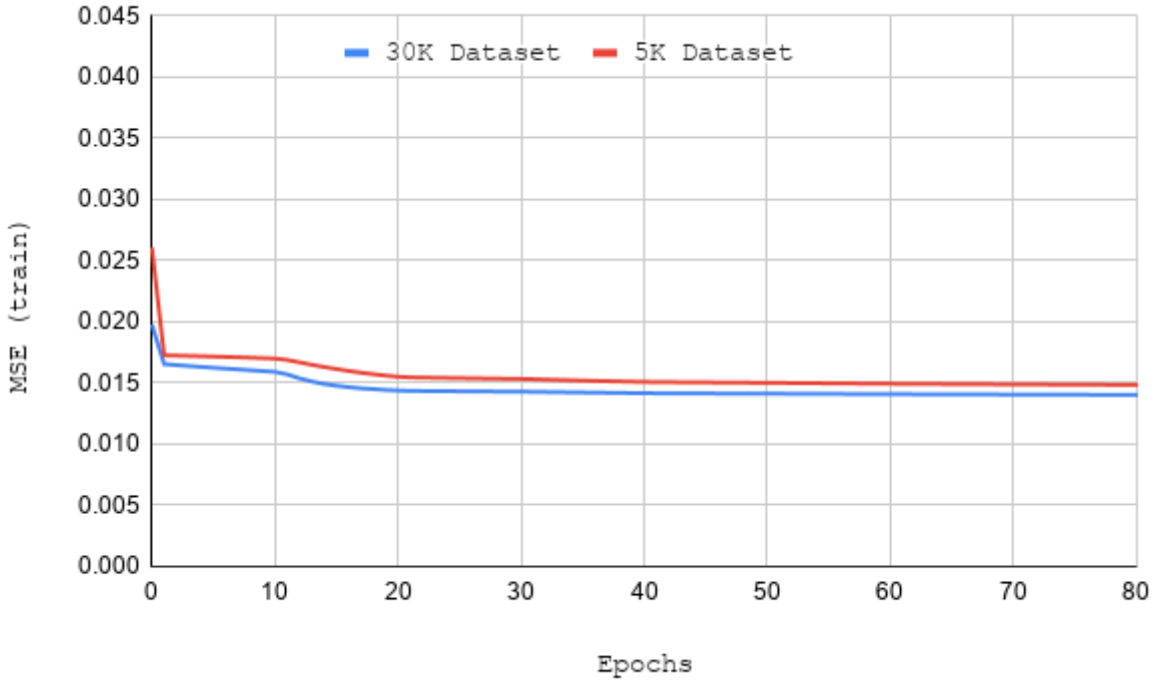

Figure 4.3: MSE loss evolution during training for DCO I architecture.

the input image to the network is of the same size as the output image from the network, meaning that the same-size image is carried throughout the network. This approach carries certain disadvantages compared to DCO II approach, which we will explain in Subsection 4.3.4.

To estimate the weights and biases of the layers in the network, DCO I implements MSE as a cost function, with a stochastic gradient descent algorithm and a momentum of 0.9. In addition, the first 20 epochs used a learning rate of 0.0001, and this rate was halved every 20 epochs to 0.00005, 0.000025, and 0.0000125, respectively, for a total of 80 epochs for the training. Training of 30K dataset took about 24 minutes per epoch for a total of 32 hours for the entire training. On the contrary, training of the 5K dataset took about 4 minutes per epoch, for a total 5 hours and 20 minutes for 80 epochs.

DCO I network contains around 667,000 trainable parameters, and to run a 180x180x3 image, we need 216MB of memory allocation. However, if the input image dimensions are 900x900x3 pixel, the memory needed for the inference is 4.0GB. Furthermore, when

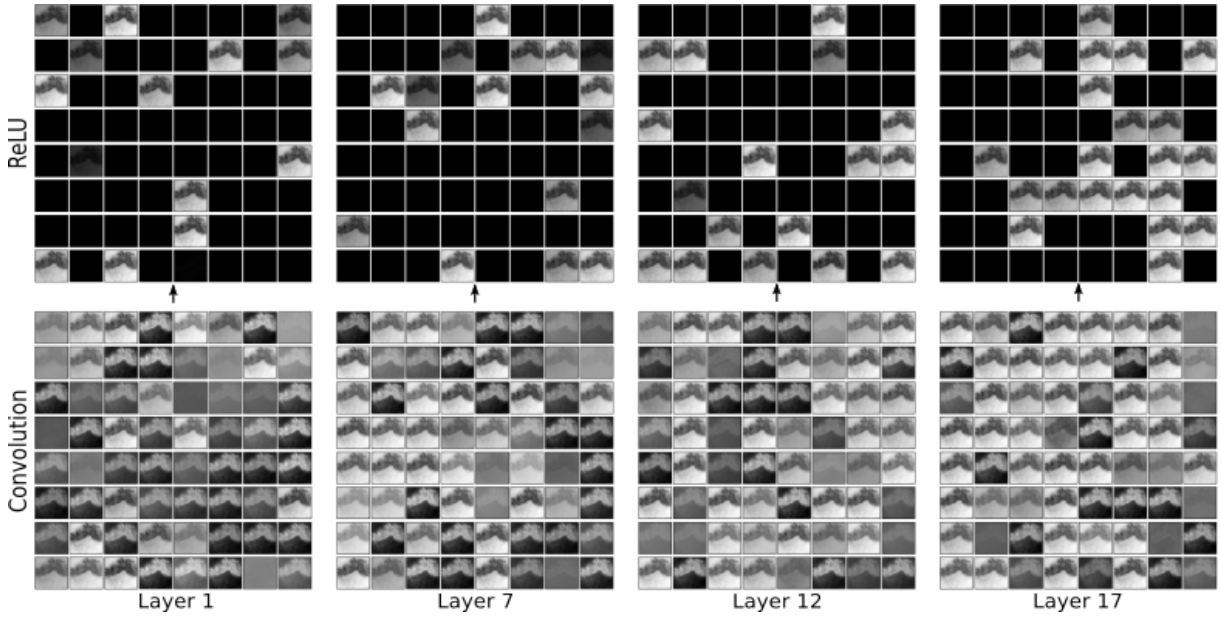

Figure 4.4: Feature maps for different layers in DCO 1 architecture. Bottom row represents a convolution for the specified layer, followed by the activation function (top row) for the same layer.

it came to inference time, 180x180x3 image took 4 seconds, while 900x900x3 image took around 12 seconds.

In Figure 4.4 we represented a feature visualization scheme for our Deep Computational Optics I architecture where each layer consists of 64 feature maps. Looking at the layer 1, we can see that the result of applying a learned filter yields different features of the skin lesion. For example, some features seem to highlight edges, while some features seem to focus on foreground (i.e., mole) and background (i.e., skin). Furthermore, looking at the activations for these features in layer 1, most of the activations are off which means the network is skipping the training for these residuals that are small. One observation to note is that the number of activations increases as the network's depth increases, suggesting that the algorithm is picking up features that represent finer details, such as high-frequency content. Looking at layers 7 and 12, it is evident that these layers exhibit features that seem to be focusing on colour and perhaps brightness judging by the higher number of features that have a lighter background than the darker features in Layer 1. One point that is not evident about this particular mole is that on the activation maps it appears to be of an upside down V-shape; however, it is actually a circular shape where the bottom

part of the mole exhibits some pigmentation that is difficult to discern from the background (i.e., skin). This complex structured pigmentation part of the mole is only picked up at the very end of the network (Layer 17), as seen on the first couple of top rows.

4.3.3 Deep Computational Optics II

DCO II was trained using the same datasets as in DCO I, and the test convergence curves for the two datasets are shown in Figure 4.5. On average, 30K dataset achieves 33.51dB PSNR, while the PSNR for the 5K dataset is approximately 32.43dB. Unlike DCO I where the training errors are symmetrical for two datasets, DCO II training errors are different. More specifically, the 30K dataset training error curves behave in a similar manner as the two curves in DCO I, where there is an error drop between epochs 1 and 20, after which the error seems to reach optimal values. This, however, is not the case for 5K dataset where the training error does not seem to reach optimal values and continues to drop throughout the whole training. This could be attributed to the fact that DCO II is a much more extensive network and requires a bigger dataset to reach optimization results faster.

DCO II design is somewhat different than the DCO I design in the sense that the HR image is down-scaled by a factor of 3 to generate a low-resolution counterpart, and as such, is carried through the network. Since the output of the network needs to be of the same size as the ground truth image, the upscaling is done at the very last layer in the network. As a result, this approach yields a smaller spatial resolution of the feature maps in DCO II architecture and improves DCO II over DCO in terms of training speed since it reduces forward/backward propagation times. In addition, it reduces inference times and memory requirements.

The estimation of network's parameters during the training phase was completed by MAE as a cost function. The optimizer for this algorithm used was ADAM optimizer with a fixed learning rate of 0.0001, no weight decay, and beta1=0.9 and beta2=0.999. Training of the 30K dataset took approximately 11.4 minutes per epoch, for a total training time of 15 hours and 12 minutes for 80 epochs. On the other hand, the 5K dataset training took 1.9 minutes per epoch, for a total of 2.5 hours for the entire training duration.

DCO II network contains over 2.7 million parameters, which is equivalent to 4 times the number of parameters contained in the previous algorithm. Despite the large number of parameters, the memory required to infer a 180x180x3 image is 84MB. However, if the input image is 900x900x3 pixel, we need 1.7GB of memory. Additionally, inference time for a 180x180x3 image is 2 seconds, while the inference time for a larger 900x900x3 image is about 6 seconds.

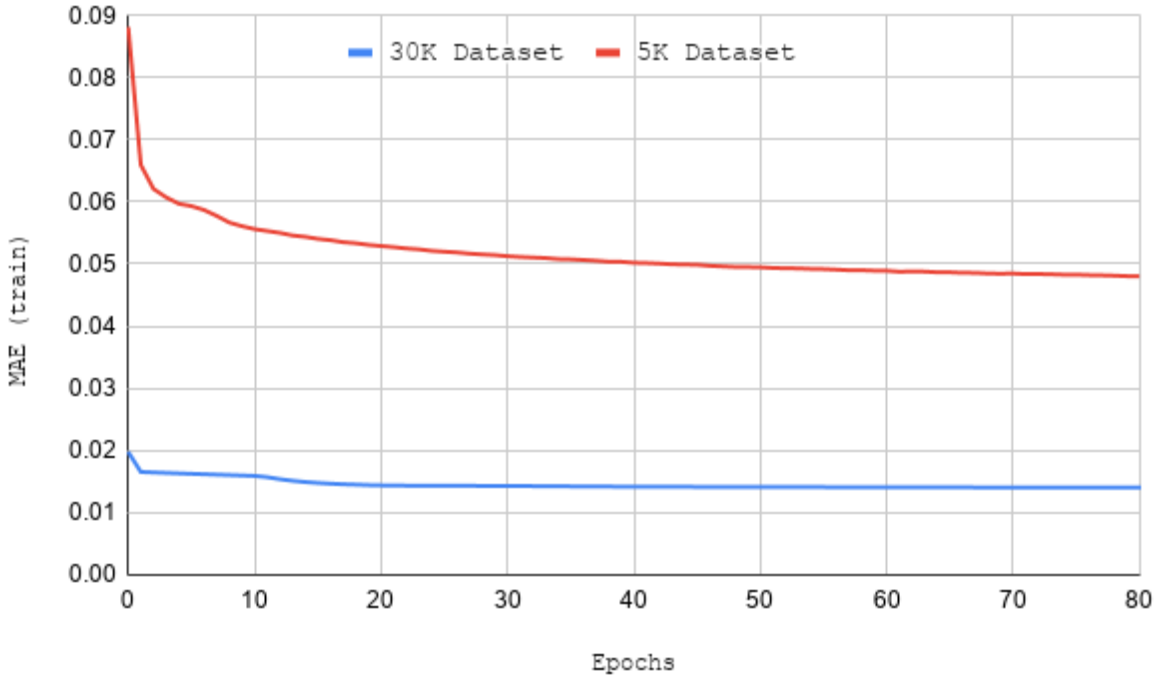

Figure 4.5: MAE loss evolution during training for DCO II architecture.

In Figure 4.6 we represented feature maps for different layers in Deep Computational Optics II architecture. Each residual block in the layer consists of 64 feature maps, followed by non-linear activation, followed by another 64 feature maps. Looking at the Layer 0, it appears that the learned filters focus on distinguishing between foreground (i.e., mole) from the background (i.e., skin), which is a similar observation that was made in the initial layers of DCO I architecture. However, looking at Layer 11, it appears that the filters are identifying all sorts of lines and edges. When it comes to Layer 21, it seems that some blobs are identified while the actual mole is barely distinguishable from the skin. Finally, looking at Layer 31, it is impossible to distinguish between the foreground and the background, and it appears that the filters are responding to very fine details in the image, suggestive of the high-frequency content.

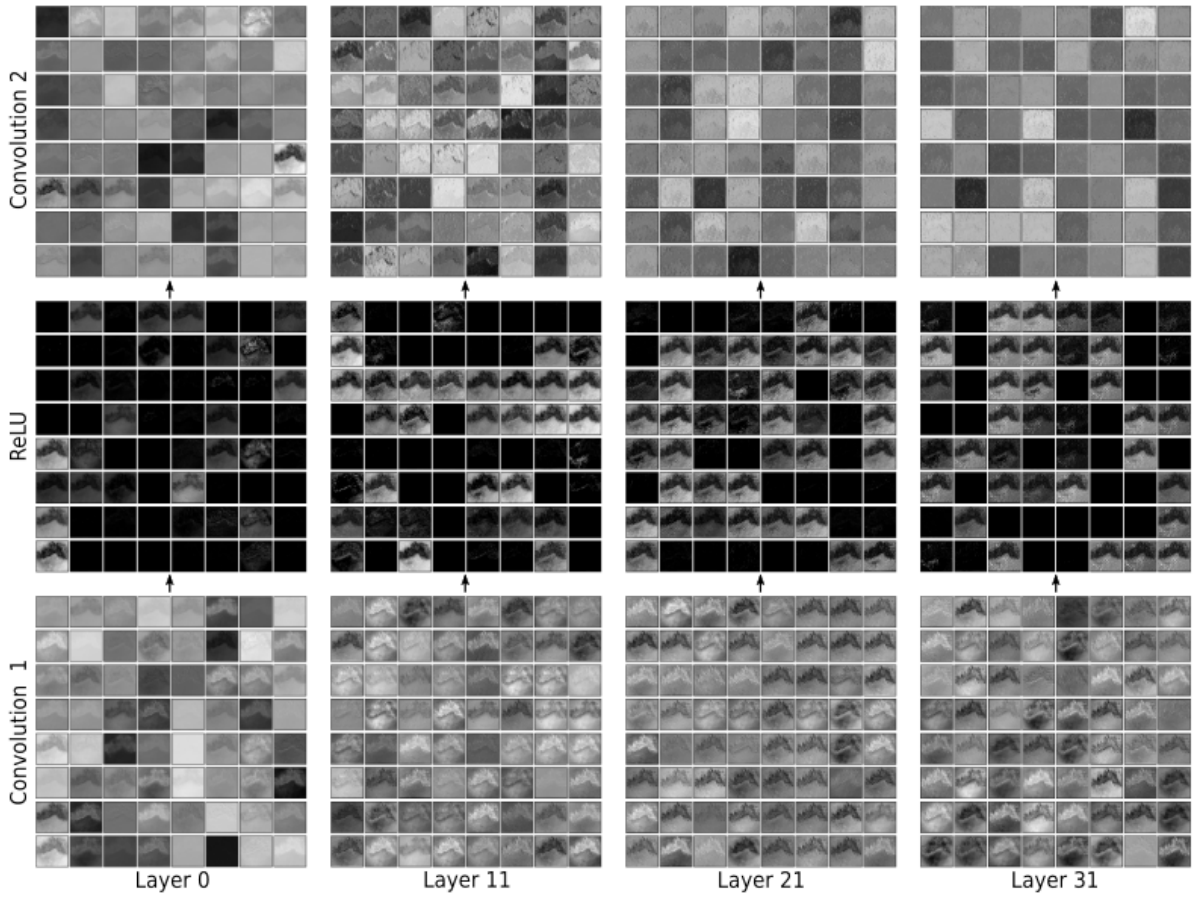

Figure 4.6: Feature maps for different layers in DCO II architecture. The bottom row is the first 64 feature maps, the middle row represents activations, and the top row is the final 64 feature maps for the specified layer.

4.4 Results

The proposed Deep Computational Optics methods were evaluated using both quantitative and qualitative approaches. Table 4.1 shows the performance metrics (PSNR and SSIM) for the bicubic interpolation method, and our two deep learning methods, DCO I and DCO II, respectively. Red text indicated the best results. This data was compiled using five different dermoscopic images captured using DELM. In terms of PSNR, DCO methods outperform the Bicubic method by a large margin, confirming that deep learning methods are a successful replacement to more conventional methods for the resolution-enhancement

	PSNR			SSIM		
	Bicubic	DCO I	DCO II	Bicubic	DCO I	DCO II
Image1	47.50	48.74	50.42	0.9407	0.9921	0.9937
Image2	38.21	39.10	39.30	0.8284	0.9306	0.9329
Image3	40.93	41.48	41.88	0.8688	0.9466	0.9493
Image4	43.25	44.21	44.85	0.9026	0.9507	0.9518
Image5	28.37	28.81	28.76	0.6923	0.8409	0.8499
Average	39.65	40.47	41.04	0.8466	0.9322	0.9355

Table 4.1: Quantitative experimental results using PSNR and SSIM, where the red text represents the best results.

problem. Furthermore, between our two deep learning methods, DCO II outperforms DCO I by 0.57dB on average, which is a large difference. This difference is even more interesting given that DCO I was trained using MSE as a loss function, which naturally favours PSNR since PSNR is derived from MSE. Furthermore, similar observations can be made for SSIM metrics, where SSIM results also confirm an absolute outperformance of deep learning methods versus the Bicubic Interpolation. Like PSNR, SSIM also reveals that DCO II method is better than DCO I method for every sample in the test results. Sample 5 is the only sample where DCO I seems to outperform DCO II by a small amount, however, SSIM results suggest otherwise. SSIM for DCO II in this case outperforms DCO I by a large margin, suggesting that PSNR is not always the most reliable metric, especially when DCO I uses MSE as a loss function for the training.

In addition to testing the efficacy of the proposed DCO methods using quantitative evaluation of resolution enhancement performance, we also applied a qualitative evaluation methods. More specifically, in Figures 4.7, 4.8, 4.9, 4.10, and 4.11 we showed (a) original (low-resolution) dermoscopic image, (b) 3X magnified image produced using Digital ELM via bicubic interpolation method, (c) 3X magnified resolution-enhanced digital ELM with DCO I method, and (d) 3X resolution-enhanced digital ELM with DCO II method. Additionally, all three methods have zoomed-in view of the same region for better visual examination purposes.

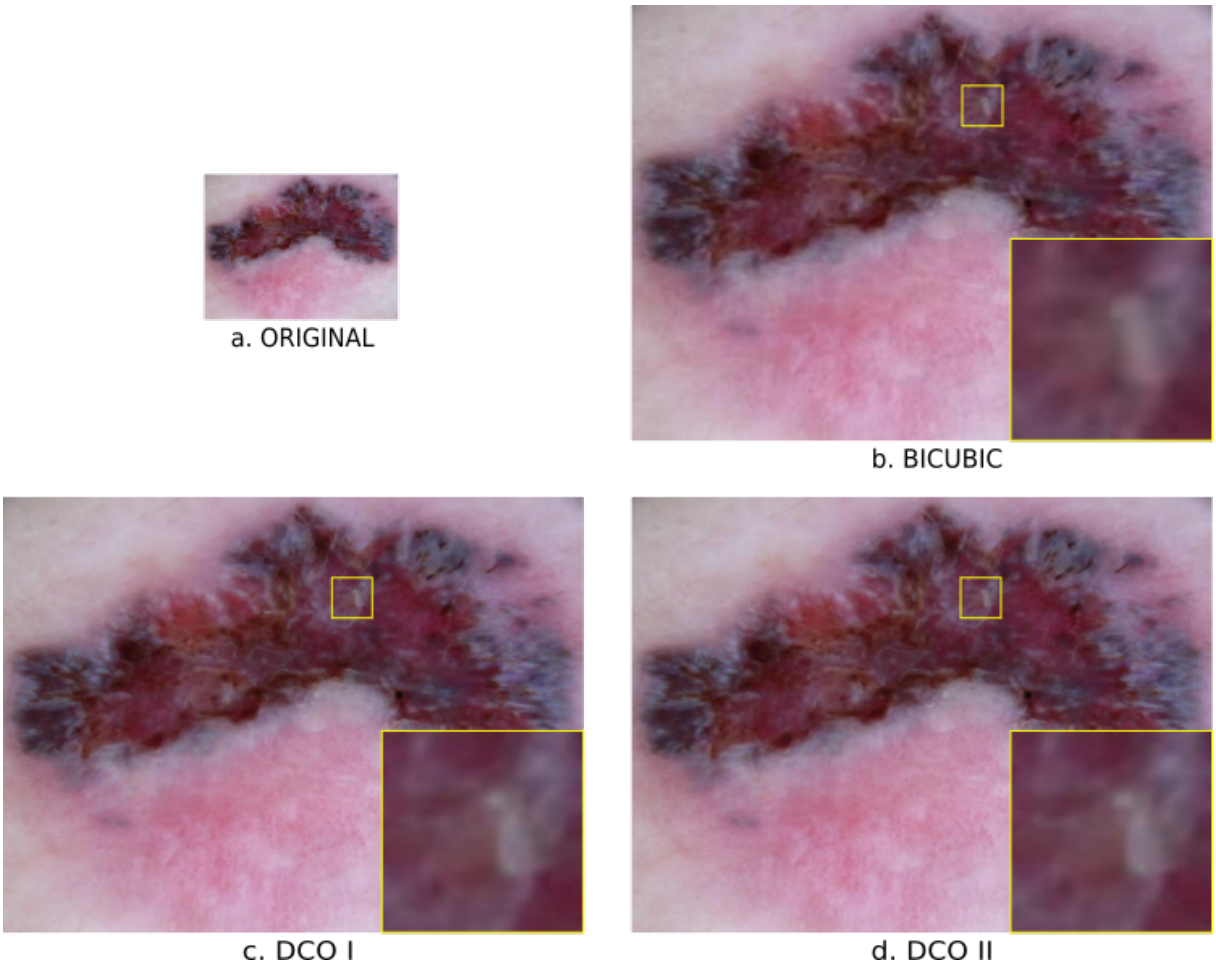

Figure 4.7: Experimental results for a skin lesion capture 1. (a) Original digital ELM (b) resolution-enhanced digital ELM obtained using Bicubic Interpolation (c) resolution-enhanced DELM via DCO I and (d) resolution-enhanced DELM via DCO II. The images produced using resolution enhanced DLM with DCO methods (c) and (d) exhibit noticeably more detail and sharpness in lesion structure when compared to the image produced with just a conventional resolution enhancement method (b). Additionally, the image produced via DCO II method seems to have somewhat better defined details when compared to the one produced with DCO I method.

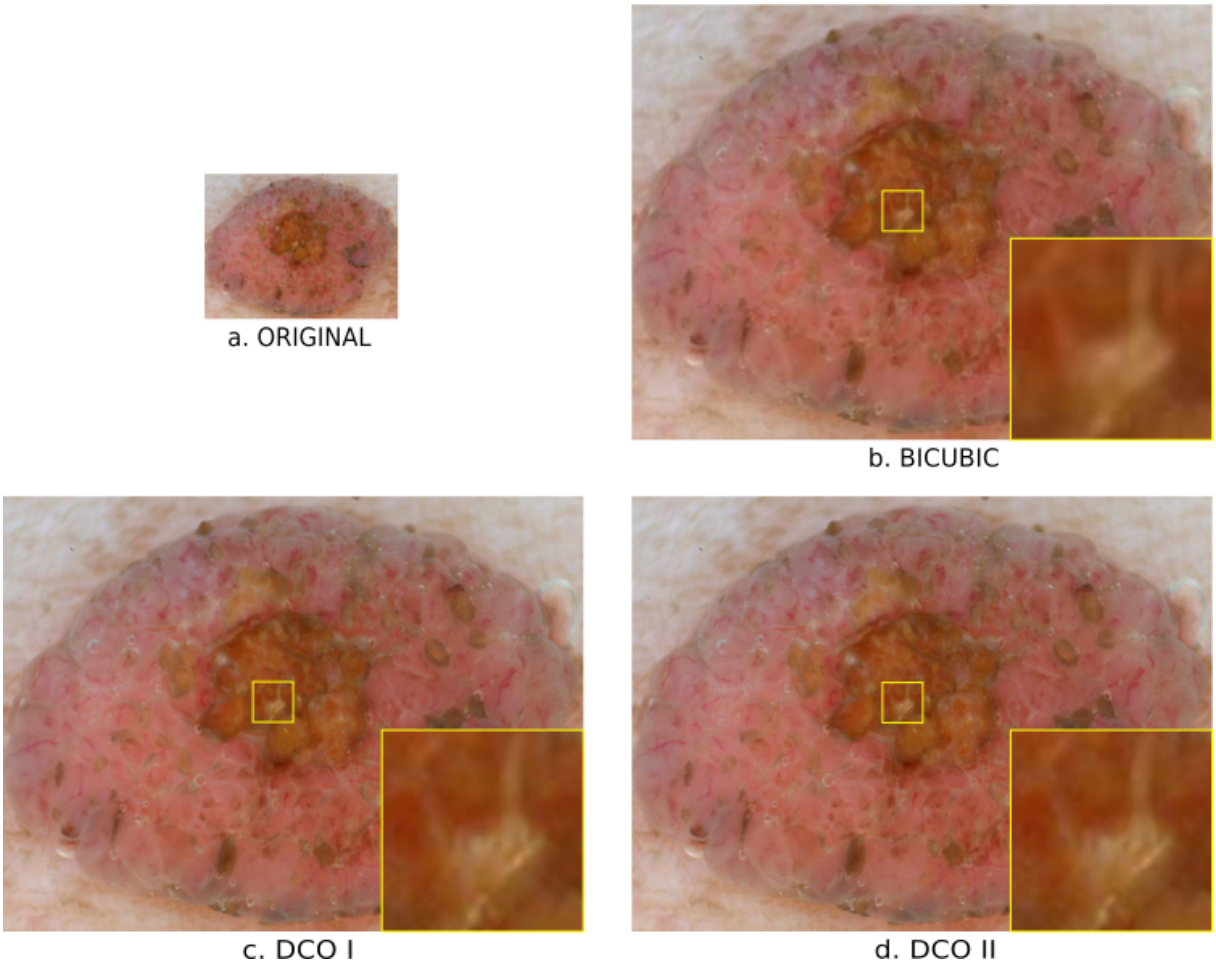

Figure 4.8: Experimental results for a skin lesion capture 2. (a) Original digital ELM (b) resolution-enhanced digital ELM obtained using Bicubic Interpolation (c) resolution-enhanced DELM via DCO I and (d) resolution-enhanced DELM via DCO II. It can be observed that out of the three methods, resolution enhanced DLM with DCO II provides the most clarity. This is specifically visible in the zoomed-in regions, where DCO II method produces the sharpest resolution image.

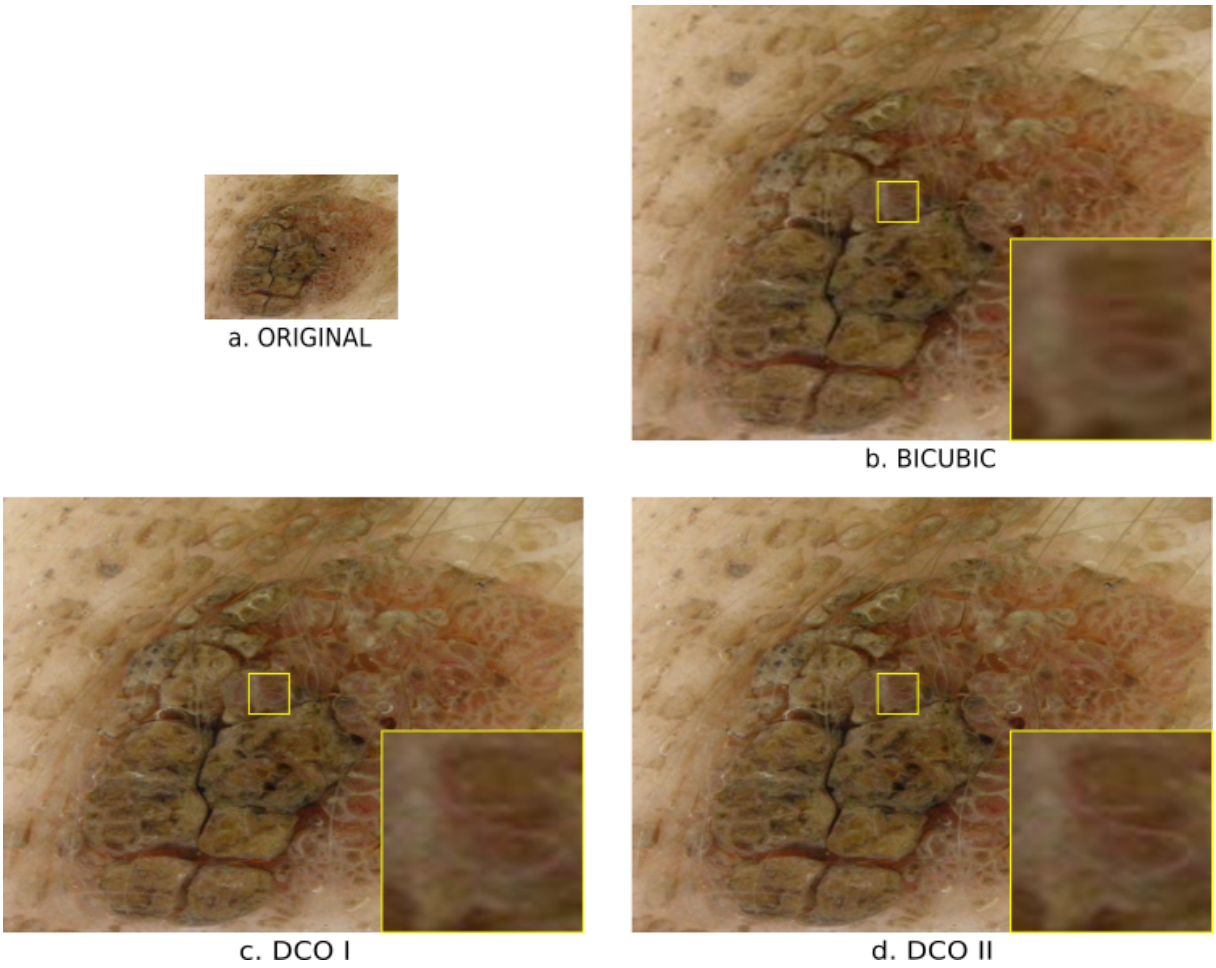

Figure 4.9: Experimental results for a skin lesion capture 3. (a) Original digital ELM (b) resolution-enhanced digital ELM obtained using Bicubic Interpolation (c) resolution-enhanced DELM via DCO I and (d) resolution-enhanced DELM via DCO II. It can be seen that the images produced using resolution enhanced DLM with DCO methods (c. and d.) exhibit noticeably more detail and sharpness in lesion structure when compared to the image produced with just a conventional resolution enhancement method (b.). Additionally, images produced via DCO methods seem to have similar characteristics. However, looking at the zoomed in regions of the two images, one can notice a shape that is somewhat reminiscent of number 3. DCO I method has the top portion of the number 3 shape blurred, while this line looks straight in DCO II.

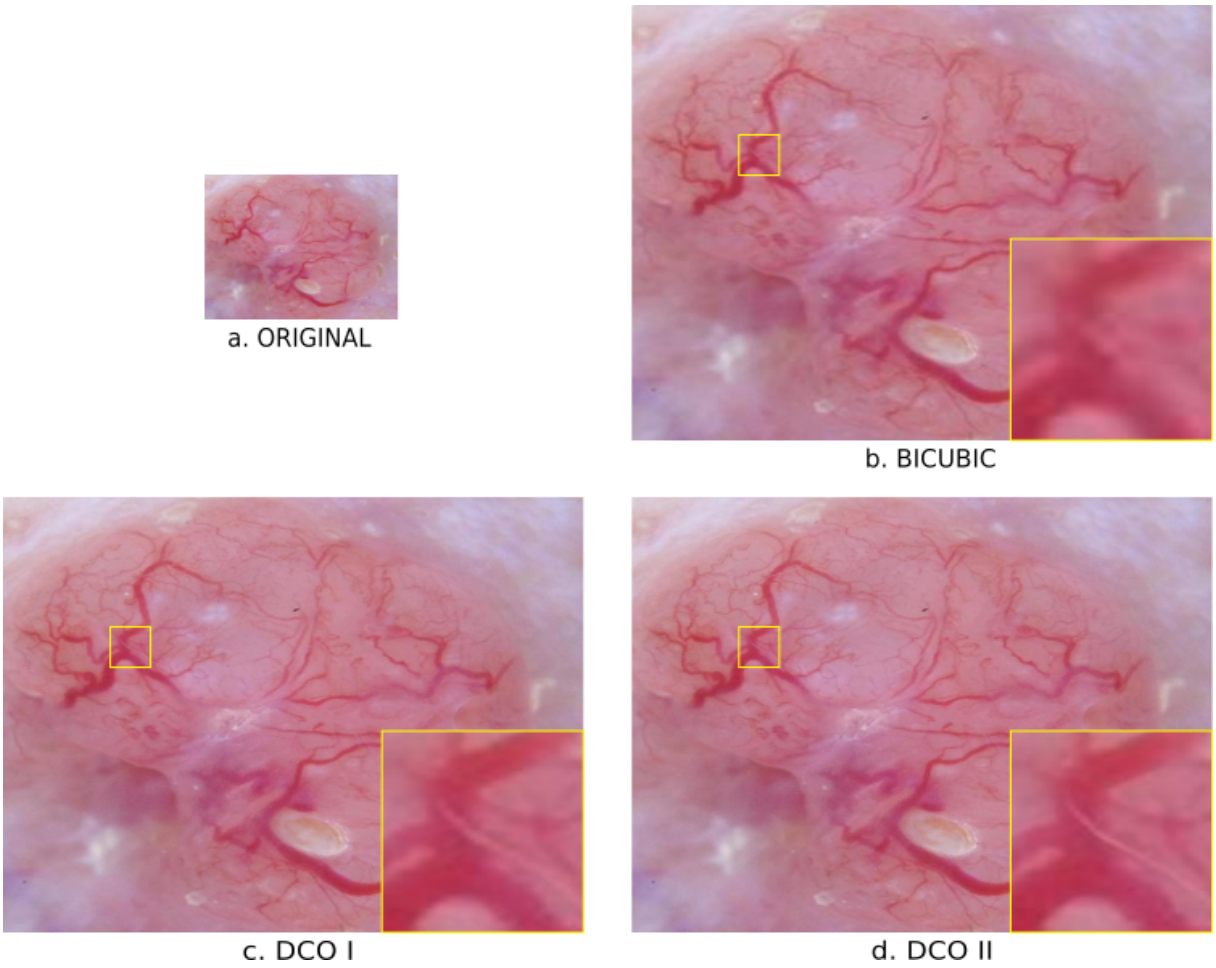

Figure 4.10: Experimental results for a skin lesion capture 4. (a) Original digital ELM (b) resolution-enhanced digital ELM obtained using Bicubic Interpolation (c) resolution-enhanced DELM via DCO I and (d) resolution-enhanced DELM via DCO II. It can be observed that the images produced using resolution enhanced DLM with DCO methods (c. and d.) exhibit noticeably more detail and sharpness in lesion vasculature when compared to the image produced with regular resolution enhancement method in b. Furthermore, zoomed-in regions reveal that the image with DCO I method is blurry when compared to the image generated using DCO II method, and this is specifically visible by observing the shape of the white diagonal line in both zoomed-in regions.

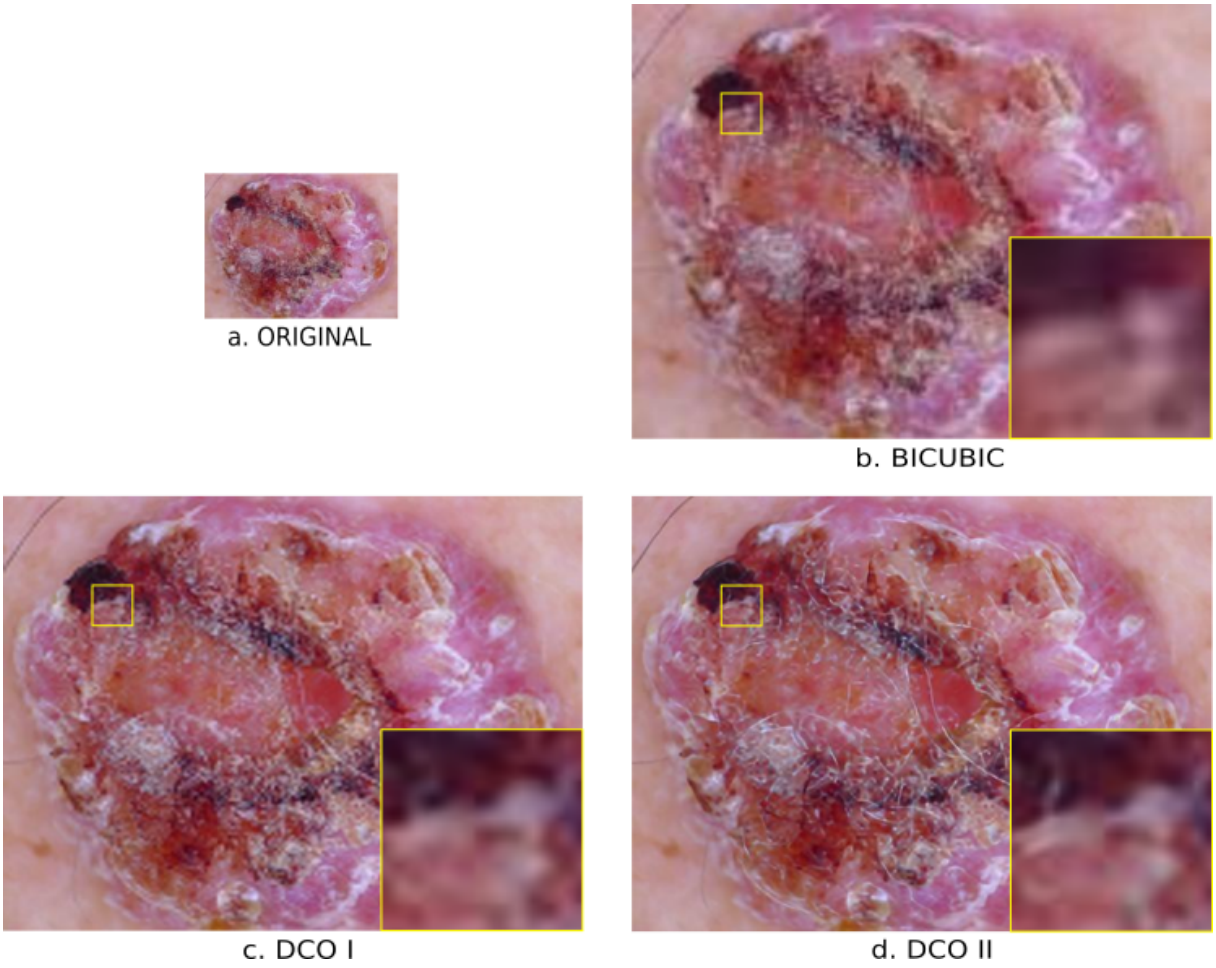

Figure 4.11: Experimental results for a skin lesion capture 5. a) Original digital ELM (b) resolution-enhanced digital ELM obtained using Bicubic Interpolation (c) resolution-enhanced DELM via DCO I and (d) resolution-enhanced DELM via DCO II. By quick examination of the images generated with different methods, it becomes evident that resolution enhanced digital ELM via DCO II is the most visually pleasing. To begin, the hair structure in the top left corner looks almost perfect with DCO II method, unlike Bicubic method where the hair structure is very blurry, and unlike DCO I method where the hair structure exhibits a pixelation artifact. Furthermore, the zoomed-in region in the DCO II image seems to resolve some of the details unlike what is observed using the other two methods.

Chapter 5

Conclusion

Conventional epiluminescence microscopes (ELM) used for skin examination are being replaced with digital ELM instruments capable of capturing and archiving images, as well as enabling an examination of lesions using computer-aided diagnosis software. A limiting factor of digital ELM instruments is the fundamental trade-off between spatial resolution and field of view (FOV), where a large FOV reduces spatial resolution, which causes a loss of fine lesion details that can be indicative of disease and malignancy. To improve the balance between spatial resolution and FOV, this thesis introduced deep computational optics (DCO), where we performed numerical magnification via a deep modeling approach to enable resolution-enhanced ELM. Experimental results using both subjective and objective image quality assessment methods revealed that DCO algorithms based on residual networks, a specialized version of deep convolutional neural network, outperformed the conventional magnification method used in practice by a large margin. Additionally, comparing our two deep computational optics algorithms, DCO II outperformed DCO I since it took half the time to train and infer one image, while it needed half of the memory allocation than DCO I. Furthermore, the DCO II method achieved better results in terms of PSNR and SSIM metrics, and produced images with more clarity, detail, and sharpness in the skin's morphological features such as structure, landscape, shape, and colour, as well as cellular layer and blood vessels, which can be essential for permitting dermatologists to better examine lesions for diagnostic purposes.

5.1 Future Work

To extend the work presented in this thesis, we suggest two topics: exploration of more efficient networks in combination with domain knowledge, and the development of a more effective loss function.

5.1.1 Network Optimization

Generally speaking, deep learning architectures outperform most of the existing algorithms mainly due to their large learning capacity, but often times they are very inefficient in their design. Even though they employ the effective universal deep learning strategies developed over the years, such as data augmentation, batch normalization, etc., it is a combination of these strategies with specific domain knowledge that yields the most successful and innovative deep learning solutions. For example, our DCO I architecture has 4X fewer trainable parameters, and yet it requires 3X the memory allocation and it is 2X slower than the DCO II architecture. It is the second architecture's realization that the input image can be low resolution without interpolation, which forces feature maps to have a significantly lower spatial resolution throughout the network with aforementioned benefits.

Additionally, a general direction for improving the resolution enhancement and image quality of digital ELM images would be to focus on alternative deep convolutional architectures. More specifically, the focus could be placed on the networks that have a larger receptive field of view, since our DCO II method with the effective FOV of 135x135 pixels outperforms DCO I method whose FOV is 41x41 mainly due to its increased receptive field. Additionally, one should take into account the memory requirement and inference speed when exploring the alternatives.

5.1.2 Loss Function

Training of a deep learning algorithm entails a continues adjustment of learnable parameters. The estimation of these parameters (weights and biases) in the neural network is accomplished with the so-called loss or cost function. As the name suggests, loss function takes an output from the network and compares it with the expected output, and treats the difference as an error or loss. In the case of the resolution enhancement problem, we have used pixel-wise loss. More specifically, the DCO I method used mean squared error (MSE), and DCO II method incorporated mean absolute error (MAE). As discussed in section 4.2, each cost function comes with its pros and cons. For example, MSE will penalize larger

errors since the difference is squared, while it will tolerate smaller errors. On the contrary, MAE will penalize smaller errors while it tolerates larger ones since it does not square the error term. For this reason, the most direct extension of the work in this thesis is to design a custom loss function that would incorporate both MSE and MAE, similar to Huber loss, where one needs to select a σ point that controls the transition from a quadratic function to an absolute value function. It would be beneficial to learn how this piece-wise function behaves during the training, and more specifically at what point during the training does it change. Perhaps, observing the σ could yield a more accurate approximation of the Huber loss. One idea unexplored in this thesis is a direct comparison of the two loss functions under the same settings. For example, one could have trained both DCO methods using MSE and MAE independently. This would perhaps reveal whether DCO methods favour one loss function over the other.

References

[1] Robert Amelard, Jeffrey Glaister, Alexander Wong, and David A Clausi. High-level intuitive features (hlifs) for intuitive skin lesion description. *IEEE Transactions on Biomedical Engineering*, 62(3):820–831, 2014.

[2] Robert Amelard, Alexander Wong, and David A Clausi. Extracting morphological high-level intuitive features for enhancing skin lesion classification. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4458–4461. IEEE, 2012.

[3] American Cancer Society. Cancer Facts & Figures 2019.

[4] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2016.

[5] Ralph Peter Braun, Harold S Rabinovitz, Margaret Oliviero, Alfred W Kopf, and Jean-Hilaire Saurat. Dermoscopy of pigmented skin lesions. *Journal of the American Academy of Dermatology*, 52(1):109–121, 2005.

[6] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[7] Huafeng Ding, Jun Q Lu, William A Wooden, Peter J Kragel, and Xin-Hua Hu. Refractive indices of human skin tissues at eight wavelengths and estimated dispersion relations between 300 and 1600 nm. *Physics in Medicine & Biology*, 51(6):1479, 2006.

[8] Claude E Duchon. Lanczos filtering in one and two dimensions. *Journal of applied meteorology*, 18(8):1016–1022, 1979.

[9] Galen C Duree Jr. *Optics for dummies*. John Wiley & Sons, 2011.

[10] Mathew Fleming. Digital dermoscopy. *Dermatologic Clinic*, 19(2):359–367, 2001.

[11] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, (2):56–65, 2002.

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[13] Leon Goldman. Some investigative studies of pigmented nevi with cutaneous microscopy. *J Invest Dermatol*, 16(6):407–27, 1951.

[14] Leon Goldman. A simple portable skin microscope for surface microscopy. *AMA archives of dermatology*, 78(2):246–247, 1958.

[15] Leon Goldman and Waldo Younker. Studies in microscopy of the surface of the skin: Preliminary report of technics. *Journal of Investigative Dermatology*, 9(1):11–16, 1947.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Dino Kabiljagic and Alexander Wong. Resolution-enhanced digital epiluminescence microscopy using deep computational optics. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII*, volume 10881, page 108811P. International Society for Optics and Photonics, 2019.

[18] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

[19] Iman Khodadad, Javad Shafiee, Alexander Wong, Farnoud Kazemzadeh, and John Arlette. Deep tissue sequencing using hypodermoscopy and augmented intelligence to analyze atypical pigmented lesions. *Journal of cutaneous medicine and surgery*, 22(6):583–590, 2018.

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[21] Harold Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[24] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[26] Zhouchen Lin and Heung-Yeung Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):83–97, 2004.

[27] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[28] Seth J Lofgreen, Kurt Ashack, Kyle A Burton, and Robert P Dellavalle. Mobile device use in dermatologic patient care. *Current Dermatology Reports*, 5(2):77–82, 2016.

[29] Ashfaq Marghoob and Ralph Braun. *An atlas of dermoscopy*. CRC Press, 2012.

[30] SW Menzies, J Emery, M Staples, S Davies, B McAvoy, J Fletcher, KR Shahid, G Reid, M Avramidis, AM Ward, et al. Impact of dermoscopy and short-term sequential digital dermoscopy imaging for the management of pigmented lesions in primary care: a sequential intervention trial. *British Journal of Dermatology*, 161(6):1270–1277, 2009.

[31] Bryan S Morse and Duane Schwartzwald. Image magnification using level-set reconstruction. 2001.

[32] Vaishali Patel and Kinjal Mistree. A review on different image interpolation techniques for image enhancement. *International Journal of Emerging Technology and Advanced Engineering*, 3(12):129–133, 2013.

[33] Hubert Pehamberger, Michael Binder, Andreas Steiner, and Klaus Wolff. In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. *Journal of Investigative Dermatology*, 100(3):S356–S362, 1993.

[34] Hubert Pehamberger, Andreas Steiner, and Klaus Wolff. In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions. *Journal of the American Academy of Dermatology*, 17(4):571–583, 1987.

[35] Domenico Piccolo, Josef Smolle, Ingrid H Wolf, Ketty Peris, Ranier Hofmann-Wellenhof, Giordana Dell'Eva, Marco Burroni, Sergio Chimenti, Helmut Kerl, and H Peter Soyer. Face-to-face diagnosis vs telediagnosis of pigmented skin tumors: a teledermoscopic study. *Archives of dermatology*, 135(12):1467–1471, 1999.

[36] Ernst A Pohle. Studies of the roentgen erythema of the human skin: I. skin capillary changes after exposure to unfiltered radiation. *Radiology*, 6(3):236–245, 1926.

[37] Ernst A Pohle. Studies of the roentgen erythema of the human skin: Ii. skin capillary changes after exposure to filtered roentgen rays and to ultra-violet radiation. *Radiology*, 8(3):185–194, 1927.

[38] June K Robinson and Brian J Nickoloff. Digital epiluminescence microscopy monitoring of high-risk patients. *Archives of dermatology*, 140(1):49–56, 2004.

[39] Mohammad Javad Shafiee and Alexander Wong. Discovery radiomics via deep multi-column radiomic sequencers for skin cancer detection. *arXiv preprint arXiv:1709.08248*, 2017.

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Hans Skvara, Ligia Teban, Manfred Fiebiger, Michael Binder, and Harald Kittler. Limitations of Dermoscopy in the Recognition of Melanoma. *JAMA Dermatology*, 141(2):155–160, 02 2005.

[43] Andreas Steiner, Hubert Pehamberger, and Klaus Wolff. In vivo epiluminescence microscopy of pigmented skin lesions. ii. diagnosis of small pigmented skin lesions and early detection of malignant melanoma. *Journal of the American Academy of Dermatology*, 17(4):584–591, 1987.

[44] W Stolz, O Braun-Falco, U Semmelmayer, and AW Kopf. History of skin surface microscopy and dermoscopy. *An Atlas of Dermoscopy*, 2004.

[45] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[46] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[47] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[49] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[50] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019.

[51] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[52] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pages 2018–2025. IEEE, 2011.

[53] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[54] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[55] Assaf Zomet and Shree K Nayar. Lensless imaging with a controllable aperture. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 339–346. IEEE, 2006.